12-2019

Truck Activity Pattern Classification Using Anonymous Mobile Sensor Data

Taslima Akter
University of Arkansas, Fayetteville

Truck Activity Pattern Classification Using Anonymous Mobile Sensor Data

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Civil Engineering

by

Taslima Akter
Bangladesh University of Engineering and Technology
Bachelor of Urban and Regional Planning, 2013
University of Toledo
Master of Arts in Geography, 2016

December 2019
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
Sarah V. Hernandez, Ph.D.
Dissertation Director

_____
Kevin D. Hall, Ph.D.
Committee Member

_____
Brent D. Williams, Ph.D.
Committee Member

_____
Kate Hyun, Ph.D.
Committee Member

ABSTRACT

To construct, operate, and maintain a transportation system that supports the efficient movement of freight, transportation agencies must understand economic drivers of freight flow. This is a challenge since freight movement data available to transportation agencies is typically void of commodity and industry information, factors that tie freight movements to underlying economic conditions. With recent advances in the resolution and availability of big data from Global Positioning Systems (GPS), it may be possible to fill this critical freight data gap. However, there is a need for methodological approaches to enable usage of this data for freight planning and operations.

To address this methodological need, we use advanced machine-learning techniques and spatial analyses to classify trucks by industry based on activity patterns derived from large streams of truck GPS data. The major components are: (1) derivation of truck activity patterns from anonymous GPS traces, (2) development of a classification model to distinguish trucks by industry, and (3) estimation of a spatio-temporal regression model to capture rerouting behavior of trucks.

First, we developed a *K*-means unsupervised clustering algorithm to find unique and representative daily activity patterns from GPS data. For a statewide GPS data sample, we are able to reduce over 300,000 daily patterns to a representative six patterns, thus enabling easier calibration and validation of the travel forecasting models that rely on detailed activity patterns. Next, we developed a Random Forest supervised machine learning model to classify truck daily activity patterns by industry served. The model predicts five distinct industry classes, i.e., farm products, manufacturing, chemicals, mining, and miscellaneous mixed, with 90% accuracy, filling a critical gap in our ability to tie truck movements to industry served. This ultimately

allows us to build travel demand forecasting models with behavioral sensitivity. Finally, we developed a spatio-temporal model to capture truck rerouting behaviors due to weather events. The ability to model re-routing behaviors allows transportation agencies to identify operational and planning solutions that mitigate the impacts of weather on truck traffic. For freight industries, the prediction of weather impacts on truck driver's route choices can inform a more accurate estimation of billable miles.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Sarah Hernandez, for her continuous guidance and inspiration throughout my Ph.D. degree. Her constructive advice and critiques enabled me to progress my research in a timely manner. I would also like extends my thanks to my dissertation committee members Dr. Kevin D. Hall, Dr. Brent D. Williams, and Dr. Kate Hyun. Their feedback during my dissertation proposal presentation helped to identify future applications of my work. I am indebted to Dr. Suman K. Mitra for his advice to estimate statistical models using spatial data and later, to publish that on a peer-reviewed journal.

The work carried out for this dissertation was supported by TRC1702 project of Arkansas Department of Transportation (ARDOT) and SPTC1501 project of Southern Plains Transportation Center (SPTC). I would like to express my gratitude for the award of funding that enabled me to undertake the research.

With sincere appreciation, I would like to acknowledge Dr. Pedro Camargo for his codes that I used to process my data. Likewise, my friends from the department of Computer Science Engineering deserve credits for their unconditional support. Especially, I would like to thank Ashutosh Dutta, Md. Monsur Hossain Tonmoy, and Kazi Abir Adnan Naval who always answered my questions about data mining and machine learning.

I would also like to acknowledge my peers from graduate school with whom I shared joyful success and sorrowful hurdles. A special thanks to Karla Diaz-Corro for the cheerful conversations including philosophy, politics, movies, and research ideas we had while traveling to conferences. Also, I would like to acknowledge the contribution of my undergraduate and graduate colleagues at the Freight Transportation Data Research Lab, particularly Rebecca

Hardwick, Chi Ngo, Ian Hargrove, and Fu Durandal, for their assistance in gathering "groundtruth" data for the model. Next, a big thanks to Sharif Mahmud who always provides me valuable writing tips and critiques my presentation styles and colors. Mariah Crews, Magdalena Asborno, and Sanjeev Bhurytal also deserve thanks for their wonderful comradeship at the lab.

Finally, I welcome this opportunity to thank my loving parents, elder sisters, younger brother, and my supportive husband for being a constant source of inspiration throughout. Without them, I would never think to complete this journey. Especially, I want to mention about my mom for her continuous care and support. Eight thousand miles and 12 hours-time difference could not stop her from being my magic wand. Every day she calls me to ask about my research progress and eases my nerve with her patient and uplifting words. Lastly, I would like to thank Thouhidul Islam for being a loving partner, best friend, and an unstinting source of support who believes in me more than I do myself.

Praise be to the Almighty for everything.

DEDICATION

To

my Mom, Dad,

Rotnapu, Rubypu, Shihab, and

Thouhid

in recognition of their inspiration, encouragement, and love.

LIST OF PUBLISHED PAPER

Chapter 4: Akter, T., Mitra, S., Hernandez, S., & Corro-Diaz, K. (2019), A Spatial Panel Regression Model to Measure the Effect of Weather Events on Freight Truck Traffic. *Transportmetrica A: Transport Science*, 2019. (Accepted for publication).

Table of Contents

List of Figures

List of Tables

Introduction

Nearly nine percent of the Gross Domestic Product (GDP) of the US economy comes from the transport of freight and thus, freight has a significant impact on the national and regional economies (Beagan, Tempesta, & Proussaloglou, 2019). The multimodal freight system moves 49 million tons of goods each day, worth more than $53 billion. Of this, trucking accounts for 69% and 64% of the market by value and weight, respectively (FHWA, 2018). The Freight Analysis Framework (FAF4), the Federal Highway Administration's (FHWA) nationwide freight demand forecasting model estimates that the weight of freight shipments moved by truck will grow 45% between 2012 and 2045 (FHWA, 2018). Hence, trucking is and will continue to be a critical component of the freight transportation system. A reliable estimation of truck travel demand is necessary for planning, design, and management of efficient freight transportation system infrastructure and operations that can accommodate the projected growth (FHWA, 2019a). Further, federal legislation including the Moving Ahead for Progress in the 21st Century (MAP-21) in 2012 and the Fixing America's Surface Transportation Act (FAST Act) in 2015 require state and Metropolitan Planning Organizations (MPOs) to consider freight in their long range transportation plans (Beagan, Tempesta, & Proussaloglou, 2019).

A goal of MAP-21 is to improve the National Freight Network to ensure efficient freight movements and economic vitality. By improving freight performance on the interstates and national highway system (NHS), MAP-21 aims to increase the accessibility of rural communities to national and international trade markets and thus, to support regional economic development (FHWA, 2019b). Similarly, the FAST Act establishes a new National Highway Freight Program (NHFP) focused on improving the efficient movement of freight on the National Highway Freight Network (NHFN) to support national economic growth (FHWA, 2017). Each state needs

to develop a State Freight Plan that addresses comprehensive freight planning activities and investments to receive funding under NHFP. Additionally, the FAST Act authorizes funds for the Intelligent Transportation System (ITS) Program. The ITS Program includes research to advance transportation safety, mobility, and environmental sustainability through electronic and information technology applications. It enhances the national freight system by supporting national freight policy goals (FHWA, 2017).

Consequently, many state and local transportation planning agencies plan to meet MAP-21 and FAST Act goals by developing policy sensitive travel demand forecasting models. Such models forecast multi-modal freight flows based on predicted origin-destination patterns, industry growth, and mode share. Traditional travel demand models are considered trip based. Trip based models first predict zonal commodity demand and supply, then predict the flow of commodities between zones, thirdly predict mode shares, and finally predict route choices. Key criticisms of trip-based models are their inability to model trip chains (Chow, Yang, & Regan, 2010). Trip chains represent the linking in time and space of consecutive trips. For example, starting from home, a truck may pick up goods, drive for several hours, drop off goods, drive a couple more hours, take a rest and fuel stop, and then drive back to their home base (Figure 1). In this example, a trip-based model would estimate four separate trips, all disconnected from each other. With that approach, it is hard to determine how policy for rest requirements, for example, may affect the order and frequency of stops in the trip chain.

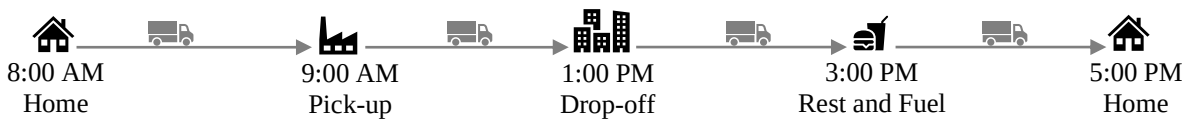


Figure 1 Example of a truck tour

In place of trip-based models, advanced freight forecasting models that incorporate policy sensitive behavioral models are increasing in popularity. However, practical implementation of advanced forecasting models that by data unavailability, specifically truck activity patterns distinguished by commodity carried or industry served. Activity patterns tied to industry served and commodity carried allow predictions of the growth/decline of certain industries to be linked to estimated truck volumes. Since truck movement data available to transportation agencies are typically void of commodity and industry information, linking truck activity to its underpinning economic drivers is a challenge. Thus, there is a need to derive truck activity patterns and tie them to the industry in ways that maintain the anonymity of the data source.

As evidenced in the FHWA's Quick Response Freight Methods (QRFM), sources of current and historical data on freight truck movements are extremely limited (Beagan, Tempesta, & Proussaloglou, 2019). Most planning agencies lack the required truck movement data needed to develop programs and policies related to infrastructure and operational solutions to mitigate bottlenecks, environmental impacts, and improve system efficiency. Public data sources such as FAF4 contain the most complete and accessible datasets to examine national trends but fail to provide data at resolutions necessary for local and regional planning. Local and regional freight studies rely on establishment surveys, travel diary surveys, roadside intercept surveys, and vehicle classification counts. These surveys may provide highly detailed information on truck movements but require expensive data collection efforts, typically provide data on a sample of the total population, and are often updated infrequently (Beagan, Tempesta, & Proussaloglou, 2019). In contrast, private sector data including transactional records, fleet operations, etc., provide necessary insights into multi-modal supply chains which can be used for freight demand

modeling but is difficult to obtain due to privacy concerns and confidentiality issues (Beagan, Tempesta, & Proussaloglou, 2019).

With recent advances in the resolution and availability of big data from cell phones and Global Positioning Systems (GPS), it may be possible to better understand freight activity patterns while overcoming the limitations presented by surveys and proprietary datasets. Mobile sensor data, from on-board or cell phone-based GPS units or Electronic Logging Devices (ELD), is increasingly available and ubiquitous. This data contains spatial-temporal position information but does not contain industry or commodity information due to privacy concerns that arise when sharing private operational data with research organizations and public sector transportation agencies. There remains a need for methodological approaches to enable the use of this data for freight planning and operations applications. **To address this methodological need, the primary objective of this dissertation is to develop spatial heuristics and machine learning algorithms to extract representative, unique, and industry specific truck activity patterns from freight big data.**

Besides planning an efficient freight transportation system through policy and infrastructure, transportation agencies are also tasked with ensuring efficient system operations during man-made and natural disasters. Specifically, adverse weather events such as tornadoes and flooding can cause significant disruptions to the freight transportation network. Such disruptions include displaced congestion effects as well as shipment delays, depreciation of goods, and inventory holding costs (Winston & Shirley, 2004) and thus, result in economic impacts to the trucking industry (Melillo, 2014). Impacts on Primary Freight Network (PFN) segments can have far reaching effects on freight movements across the nation. For example, Ivanov et al. (2008) estimated that due to two corridor closures in Washington caused by storm

events, the total loss from freight delay was almost $75 million. Transportation agencies need to understand the effects of weather events on truck traffic patterns if they are to propose and implement winter maintenance programs, alternative routes, emergency management operations, and identify critical network links. Considering that truck drivers follow strict delivery schedules, drivers may not be able to cancel and/or postpone their trips to avoid adverse weather but instead may choose an alternate route. **To support transportation planning agencies, another objective of this dissertation is to develop a predictive model that captures the spatial and temporal rerouting behavior of freight trucks due to adverse weather events.**

The approaches presented in this dissertation allow anonymous GPS data to be linked to the industry served and commodity carried without violating privacy concerns. Ultimately, the methods to tie truck movement data to industry and commodity, close the identified research gap and open the door for the development of advanced freight forecasting models such as Activity Based Models (ABMs). Federal, state, and local transportation agencies can use industry-specified truck activity patterns for development, calibration, and validation of advanced freight forecasting models, ultimately allowing them to satisfy MAP-21 and FAST Act requirements.

Four models were developed in this dissertation: (1) a multinomial logistic regression model to identify truck operational characteristics that differ by commodity to support development of feature extraction algorithms, (2) a *K*-means clustering model to extract representative freight activity patterns that can support and validate activity-based models, (3) a random forest model to classify daily activity patterns by freight industry that can be used in commodity-based freight forecasts, and (4) a spatio-temporal regression model to capture rerouting behaviors of truck drivers due to extreme weather events to better plan adverse weather management and operations. To evaluate truck re-routing behaviors, we combined truck volume

data from a fixed sensor network (e.g., Weigh-in-Motion sensors) with weather data from the atmospheric data assimilation system (e.g., Modern-Era Retrospective analysis for Research and Applications, Version 2).

This dissertation is organized as follows. Chapter 1 describes the development of algorithms for pre-processing GPS data. This includes descriptions of the data quality control, stop identification, path identification, and trip identification process. This chapter also presents the derivation of operational characteristics from the pre-processed GPS data and a multinomial logistic regression model that identifies commodity specific freight operational characteristics. Chapter 2 presents an unsupervised learning algorithm, *K*-means clustering, to extract unique and representative daily activity patterns from operational characteristics derived from GPS data. Chapter 3 presents a supervised learning, random forest model to predict industry served based on operational characteristics derived from GPS data. Chapter 4 presents a spatio-temporal model to capture the rerouting behaviors of freight trucks during adverse weather conditions. Chapter 5 presents the applications of the developed models including commodity flows on roads, truck load distribution on pavements, and changes in Vehicle Miles Traveled (VMT) due to weather events. The dissertation concludes by highlighting significant findings, noting limitations, and suggesting future improvements.

References

Beagan, D., Tempesta, D., & Proussaloglou, K. (2019). *Quick Response Freight Methods*. Retrieved from https://ops.fhwa.dot.gov/publications/fhwahop19057/fhwahop19057.pdf.

Chow, J., Yang, C., & Regan, A. (2010). State-of-the Art of Freight Forecast Modeling: Lessons Learned and the Road Ahead. *Transportation, 37*(6), 1011-1030. doi:10.1007/s11116-010-9281-1.

FHWA. (2017). Fixing America's Surface Transportation Act or "FAST Act". Retrieved from https://www.fhwa.dot.gov/fastact/summary.cfm.

FHWA. (2018). Status of the Nation's Highways, Bridges, and Transit Conditions and Performance: 23rd Edition: Part III: Highway Freight Transportation - Report to Congress. Retrieved from https://ops.fhwa.dot.gov/freight/infrastructure/nfn/rptc/cp23hwyfreight/iii_ch11.htm.

FHWA. (2019a). Freight Economy. Retrieved from https://www.fhwa.dot.gov/freighteconomy/.

FHWA. (2019b). Moving Ahead for Progress in the 21st Century act (MAP-21). Retrieved from https://www.fhwa.dot.gov/map21/summaryinfo.cfm.

Ivanov, B., Xu, G., Buell, T., Moore, D., Austin, B., & Wang, Y. (2008). *Storm Related Closures of I-5 and I-90: Freight Transportation Economic Impact Assessment Report, Winter 2007-2008.* (No. WA-RD 708.1). Retrieved from https://rosap.ntl.bts.gov/view/dot/17218.

Melillo, J. M. (2014). Climate Change Impacts in the United States. *Third National Climate Assessment, 52*. Washington, DC: U.S. Global Change Research Program. Retrieved from https://doi.org/10.7930/J0Z31WJ2.

Winston, C., & Shirley, C. (2004). The Impact of Congestion on Shippers' Inventory Costs. *Federal Highway Administration, Washington DC.*

Chapter 1

1 Freight Operational Characteristics Mined from Anonymous Mobile Sensor Data

1.1 Abstract

Effective Transportation Performance Measurement (TPM) benefits from ubiquitous system coverage. In the context of freight oriented TPM, traditional performance monitoring devices like inductive loops, cameras, manual counts, etc., may fail to provide comprehensive and high resolution coverage, e.g., providing only volume counts with no indication of trip linkages typically for a small subset of links across a large network. New sources of big data from mobile sensors including on-board Global Positioning System (GPS) devices allow more universal network coverage and insights into trip chaining behaviors. However, to gain actionable insights into system performance from large and noisy streams of mobile sensor data, it is necessary to mine it for relevant operational characteristics of the trucks it represents. Such characteristics include stop locations, stop duration, stop time of day, trip length, and trip duration. To address this methodological need, we developed three heuristic algorithms, i.e., *stop-identification*, *path-identification,* and *trip identification*. To address the issue of determining relevant operational characteristics, we developed a Multinomial Logistic (MNL) regression model. We interpret relevancy as the ability of each operational characteristic to predict commodity carried, which is removed from GPS data to protect privacy, e.g., anonymized. The MNL model relates operational characteristics to commodity carried which is a critical data gap that currently limits development of advanced freight forecasting models.

1.2 Introduction

Effective Transportation Performance Measurement (TPM) benefits from ubiquitous system coverage. Due to the significant impact of trucking on the economy, infrastructure, and environment, it is essential that transportation agencies consider freight movements in TPM. To ensure freight needs are met, federal legislation (e.g., the Fixing America's Surface Transportation or FAST act), mandates a process of selecting performance measures, setting performance targets, and establishing a freight plan that aligns with the broad goal of improving the National Highway Freight Network to ensure economic competitiveness.

With the push toward more accurate and detailed freight performance measurement and system planning there is an ever-increasing need to better understand and measure freight truck movements at high levels of temporal and spatial disaggregation (Roorda, Cavalcante, McCabe, & Kwan, 2010). In the context of freight oriented TPM, traditional performance monitoring devices like inductive loops, cameras, manual counts, etc., may fail to provide comprehensive, high-resolution coverage of the transportation network. For instance, static devices like loops and cameras only provide data for the link on which they are located and typically measure only volume with no indication of trip linkages. Acquiring the data needed for system wide TPM is a challenge for transportation agencies and a special challenge if freight data is needed. Since freight operations are carried out primarily by private entities, e.g., shippers, carriers, businesses who collect significant data on their operations, this data is often not made readily available due to privacy concerns.

New sources of big data from mobile sensors including cell phones and Global Positioning System (GPS) devices allow more universal network coverage and insights into trip chaining behaviors. Recently, carrier collectives have made available large streams of

anonymized Global Positioning System (GPS) data (CPCS, 2018). This GPS data typically contains the timestamp, latitude and longitude position (e.g., ping), and point speed data for a sample of trucks operated by major freight carriers. All data regarding the carrier, fleet operator, driver, cargo/commodity, and trip purpose are removed from the data to protect privacy. Therefore, the anonymized data must be mined to extract relevant data for planning applications such as stop location/purpose, trip purpose, and commodity carried. Moreover, data mining should not reveal private information such as company/fleet identification or name.

Freight activity insights derived from truck GPS data have been applied in practice to support a variety of freight planning efforts including: freight forecasting tools like activity-based and truck touring models (Bassok, McCormack, Outwater, & Ta, 2011; Kuppam et al., 2014; Camargo, Hong, & Livshits, 2017), estimating origin-destination truck flows (Zanjani et al., 2015; Sharman & Roorda, 2011), improving the estimation of freight performance measures (Liao, 2009; Ma, McCormack, & Wang, 2011), and ranking roadway bottlenecks (Zhao, McCormack, Dailey, & Scharnhorst, 2013). Although these studies used truck GPS data to develop and/or validate their forecasting models, they fall short in identifying underlying relationships between truck activity and commodity carried. Such a relationship is key in forecasting models that make use of economic forecasts.

For long-haul trips, average trip length (ATL) varies by commodity carried (Beagan, Tempesta, & Proussaloglou, 2019; Evans, Kassinger, Cooper, & Kincannon, 2004). However, ATL is the only trip characteristics available from most surveys like the Vehicle Inventory and Use Survey (VIUS) used for freight analysis and it is likely other trip characteristics that vary by commodity (Evans, Kassinger, Cooper, & Kincannon, 2004). Unfortunately, being a national inventory conducted annually, VIUS does not tell us about daily trip patterns, trip chains, or

shorter trips resulting from needs for rest breaks, fuel, etc. Therefore, it is necessary to identify key freight operational characteristics from a state level data that can be used in comprehensive freight planning.

To address the critical need for methods to extract operational characteristics from mobile sensors data, we present three transferable heuristic algorithms to identify stop characteristics and trip characteristics from truck GPS data: (1) *stop-identification* to aggregate pings (latitude, longitude, timestamp data points) into freight activity stops, i.e., pick-up/drop-off or rest stops, (2) *path-identification* to convert sparse pings into complete, fully connected paths on a dense transportation network, and (3) *trip identification* to extract operational characteristics by combining results of *stop identification* and *path-identification* algorithm. The algorithms were applied to a sample of 338 million GPS pings collected from major trucking companies and cover a statewide region. Finally, to identify the operational characteristics that can be linked to commodity carried, we developed a multinomial logistic regression (MNL) model. Application of these approaches to mobile sensor data enables such sources of big data to be used effectively for TPM.

1.3 Background

This section reviews prior research focused on heuristic approaches, methods, and models that were used to extract freight operational characteristics from large streams of truck GPS data.

1.3.1 Stop identification

The premise of stop identification is to determine the locations of potential activity stops (e.g., fuel stops, rest stops, and pick-up/delivery for freight trucks) within large streams of GPS pings. Simple algorithms consider a stop to be the location where the vehicle's instantaneous

speed is recorded as zero. Minor modifications may assume the speed to be below a given threshold, say 3 miles per hour (mph). However, these simple approaches may overestimate the number of stops made by a vehicle by not grouping consecutive (redundant) pings representing zero or low speed into a single stop. In short, effective algorithms should combine consecutive, low speed pings into clusters, determine the physical location of the stop within the cluster, and calculate a stop duration considering all pings in the cluster.

Existing stop identification algorithms used geographic bounding boxes and rule-based approaches to define stop clusters (Greaves & Figliozzi, 2008; McCormack, Ma, Klocow, Currarei, & Wright, 2010; Thakur et al., 2015; Camargo, Hong, & Livshits, 2017). Greaves and Figliozzi (2008) developed a stop identification algorithm for commercial vehicles and used the time difference between GPS-to-satellite communications to determine if the vehicle was stopped. The algorithm considered a time threshold of 4 minutes (240 seconds) and a geographic distance threshold of around 20 feet (6 meters) to identify a stop. If a vehicle repositioned by less than the defined threshold, regardless of the time elapsed, they performed a manual inspection to check whether it was a short stop. However, relying on manual inspections is time-consuming for a large dataset. McCormack, Ma, Klocow, Currarei, and Wright (2010) identified delivery stops by defining a threshold of 3 minutes (180 seconds) for dwell time (i.e., duration of a vehicle's engine as off or idle status). To avoid redundant GPS pings of an idle truck, their algorithm removed data points where the distance between two consecutive pings was less than 65 feet (about 20 meters). Though this filtered out false trips, it removed data that could be significant for deriving freight operational characteristics like service times (i.e., the time for a truck to unload and start the next trip).

The stop identification algorithm developed by Camargo, Hong, and Livshits (2017) expanded on the abovementioned research by using coverage and space mean speed in addition to dwell time to define a stop. After grouping pings for which the travel speed between consecutive GPS records was less than 5 mph (8 km/h), they assessed the coverage of the set of pings. If a truck traveled less than 0.5 miles (about 800 meters) between stops, pings were combined to represent a single stop. The geometric center of the stop cluster was defined as the stop location. The stop identification method developed by Camargo, Hong, and Livshits (2017) was used in this work several modifications to ensure transferability among datasets, e.g. metropolitan vs statewide scales.

1.3.2 Path identification

Path identification, also known as map-matching, refers to the process of identifying the network link that corresponds to each GPS ping (a latitude, longitude, and timestamp data triples). Existing map-matching algorithms were developed based on the premise of assigning the pings to their closest network link and then connecting disparate links via shortest path finding algorithms (Giovannini, 2011; Quddus & Washington, 2015; Camargo, Hong, & Livshits, 2017). Giovannini's (2011) algorithm re-constructed routes from low-sample rate GPS data, e.g., around one mile between pings, using a Bayesian approach. Quddus and Washington (2015) developed a weight-based shortest path and vehicle trajectory aided map-matching algorithm to determine the network link corresponding to each GPS ping based on proximity, among other factors, for a sparse road network.

With temporally sparse GPS data simple matching of the GPS ping to the closest link may not result in a complete and connected path. For example, many network links may be traversed between consecutive pings if the pings are recorded only every 15 minutes and a

vehicle is traveling at highway speeds of 55 mph, and thus there would be gaps when constructing the complete path of the truck from origin to destination. Camargo, Hong, and Livshits (2017) addressed this gap by determining a fully connected complete path between sparse pings by applying shortest path algorithms. The map-matching algorithm developed by Camargo, Hong, and Livshits (2017) was used in this paper with several modifications to ensure route accuracy for a denser road network.

1.3.3 Freight operational characteristics from mobile sensor data

Identifying stops and routes from GPS data allows us to compute network volumes, link/corridor speeds, identify bottlenecks, and estimate many other performance metrics for TPM. For freight oriented TPM, it is also important to differentiate performance measures by operational characteristics like trip type (e.g., long-haul and short-haul trip), stop and trip purpose (e.g., rest, pick-up delivery, pass through), and industry served to enhance our understanding of economic impacts tied to freight movements.

Yang, Sun, Ban, and Holguín-Veras (2014) characterized freight delivery stops from other types of stops using GPS data and a Support Vector Machine (SVM) method. Three parameters, e.g., stop duration, the distance to the center of the city, and the binary distance to a stop's closest bottleneck, served as input features of the SVM and produced minimal error of 0.2% (Yang, Sun, Ban, & Holguín-Veras, 2014). Based on trip length and number of trips derived from truck GPS data, Zanjani et al. (2015) distinguished light duty local delivery trucks from long haul operations using heuristic approaches. A local delivery truck was characterized as making more than five trips per day, none more than 100 miles in length. In combination with a driver survey, Jing (2018) analyzed stop purpose, stop duration, and stop time of day. Her study found four types of overnight, urban truck tours, i.e., one pickup followed by one delivery,

multiple consecutive pickups followed by one delivery, one pickup followed by multiple consecutive deliveries, and multiple consecutive pickups followed by consecutive deliveries.

None of the studies mentioned above were aimed at identifying or deriving freight operational characteristics that distinguish freight daily activity patterns by commodity carried or industry served. Knowledge of industry served can be used to estimate economic impacts associated with performance measurements, prioritize critical freight corridors according to key industries, and relate changes in economic conditions to transportation system performance. This paper relates operational characteristics defined from stop and path identification algorithms to trip type, stop and trip purpose, industry associated trip chaining, or activity patterns.

1.4 Methodology

The methodology consists of four key approaches: (1) establishing consistency and relevancy of GPS data to improve algorithm performance, (2) modification of stop and path identification algorithms, (3) derivation of truck operational characteristics, and (4) development of a multinomial logistic (MNL) regression model.

1.4.1 Data consistency and relevancy

Most commonly used truck GPS data sources require pre-processing to remove noise and other inconsistencies (Camargo, Hong, & Livshits, 2017). Hence, we developed an algorithmic data validation approach to improve data consistency and relevancy. The approach identifies a complete truck record for input into the stop identification and path identification algorithms. Complete truck records were defined as those that represented an over the road truck movement with logical start and end positions, speeds, and accelerations.

The *consistency and relevancy* (CR) algorithm identified the inconsistent truck trajectories and flagged those records for further analysis (Figure 1.1). First,

acceleration/deceleration rate of each truck for each pair of consecutive pings was calculated and pings that produced acceleration/deceleration rates above a predefined threshold of 2.24 ft/s^2, corresponding to 85th percentile average acceleration rate of heavy trucks were removed (Pline, 1999). Next, the total number of pings corresponding to each truck record was calculated and truck records that had fewer pings than the threshold count ($\boldsymbol{p_{count}}$) were removed. Then, the space-mean-speed and travel time between each consecutive pair of pings were calculated. Truck records were removed when the calculated space-mean-speed (SMS) exceeded the speed limit ($\boldsymbol{s_{max}}$) for a threshold time ($\boldsymbol{t_{max}}$). Lastly, the geographic coverage area for each truck was calculated and any truck records that had a smaller geographic coverage area than the threshold area ($\boldsymbol{c_{max}}$) was removed. Geographic coverage was defined as the diagonal of the rectangular bounding box that surrounds all pings of a truck.

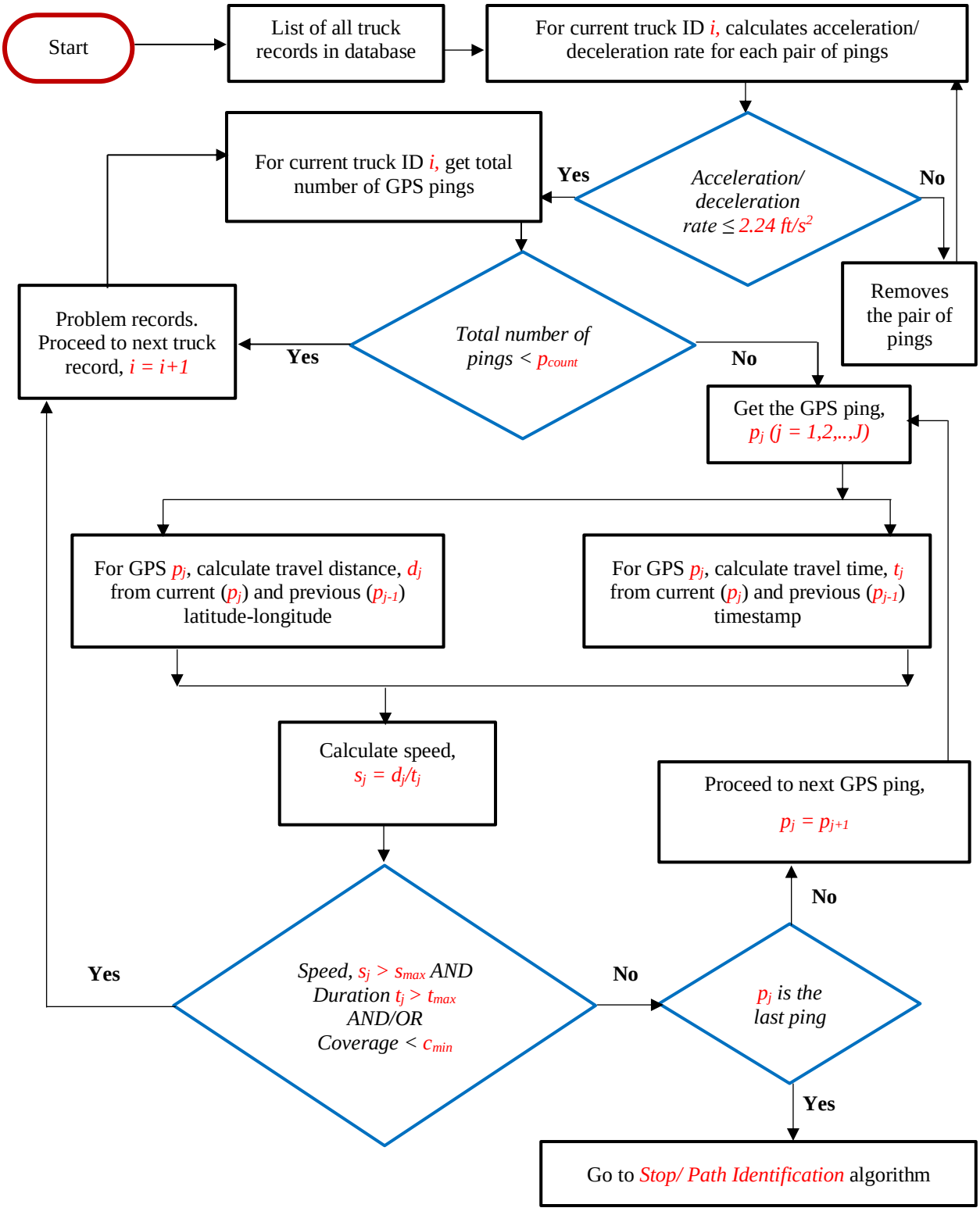


Figure 1.1 *Consistency* and *relevancy* algorithm

1.4.2 Stop and path identification algorithms

The *stop identification* algorithm developed in this paper was modified from Camargo, Hong, and Livshits (2017). We extracted stops from the set of valid truck records identified through the CR algorithm. The stop identification algorithm calculated the space-mean speed (s_j) between consecutive pings (p_{j-1} and p_j) (Figure 1.3). If the space-mean-speed was less than a defined threshold speed (s_{min}) (i.e., 3 mph were used in this paper) for at least threshold time (t_{min}) (i.e., 5 minutes were used in this paper), the algorithm continued by calculating the speed between the next pair of consecutive pings. Next, a series of the pings that passed the speed and time criteria, $\{p_j, p_{j+1},\ldots,p_J \mid s_j \leq s_{min\ AND}\ t_j \geq t_{min}\}$ were collected. Following, the total stop coverage (c_T) and the total stop duration (t_{TQ}) for all consecutive pings from the series were calculated (Eq. 1.1 and 1.3). If the total coverage for the series of the pings was less than c_{max} (i.e., 0.2 miles was used in this paper), then the series was considered as a stop-cluster (Q) (Eq. 1.2). Although Camargo, Hong, and Livshits (2017) specified the geographical center of the stop-cluster (Q) as the stop location of the cluster, we noticed that the geographical center could be incorrect occasionally (e.g., in the middle of a road). Hence, we used the first identified stop's location (l_j) as the stop location for the stop cluster (Q). Ultimately, we identified a set of stop locations (i.e., pick-up/delivery stops, rest or fuel stops) along with stop time of day, stop duration, and stop coverage for each truck record.

$$c_T = \text{geographical coverage of all consecutive stops} \tag{1.1}$$

$$Q = \{p_j, p_{j+1}, .., p_J \mid c_T < c_{max}\} \tag{1.2}$$

$$t_{TQ} = \sum_{j=1}^{J} t_j \quad ; \forall j \in Q \tag{1.3}$$

Where,

$\boldsymbol{c_T}$ = Diagonal of the rectangular bounding box that surrounds all consecutive stops

$\boldsymbol{Q}$ = A stop cluster of consecutive stop pings

$\boldsymbol{p_j}$ = GPS pings, where $j = 1, ..., J$

$\boldsymbol{t_j}$ = Calculated travel time from current (p_j) and previous (p_{j-1}) timestamp, where $j = 1, ..., J$

$\boldsymbol{t_{TQ}}$ = Total stop duration for a series of consecutive stops, Q

The *path identification* algorithm identified the set of links that comprised the complete path between consecutive pings (Figure 1.4). First, a spatial buffer ($\boldsymbol{b}$) was created around each network link ($\boldsymbol{r_l}$). Next, each GPS ping ($\boldsymbol{p_j}$) was paired with a network link based on proximity. The link buffer helped to account for small, inherent inaccuracies in the GPS ping positions. After associating each ping ($\boldsymbol{p_j}$) with a link ($\boldsymbol{r_l}$), it is possible that the set of links comprising the path were not fully connected. This was due to the temporal sparsity of the GPS ping data. To repair this gap in the path, the shortest path between consecutive pings was determined (Figure 1.2). The link cost (i.e., travel time was used in this study) calculation for using those routes was shown in Eq. 1.4, 1.5, and 1.6. Thus, we estimated a complete but shortest path for each truck. Due to the temporal coarseness of the GPS pings and the density of the network links, this was a critical step in determining, at the aggregate level, the volume of trucks along each link in the network and, at the disaggregate level, the accurate distance and travel time for each truck record.

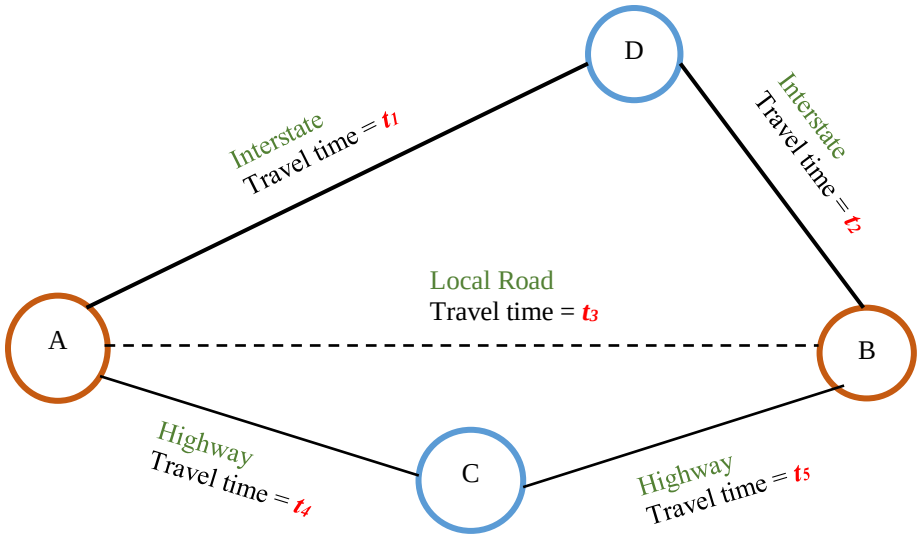


Figure 1.2 Shortest path considering travel times

1st Alternative: A → D → B	$\mathbf{L_i} = t_1 + \mathbf{t_2}$	(1.4)
2nd Alternative: A → B	$\mathbf{L_l} = t_3$	(1.5)
3rd Alternative: A → C → B	$\mathbf{L_h} = t_4 + \mathbf{t_5}$	(1.6)

Where,

$\mathbf{L_i}$ = Link cost for path 1 using interstates

$\mathbf{L_l}$ = Link cost for path 2 using local roads

$\mathbf{L_h}$ = Link cost for path 3 using highways

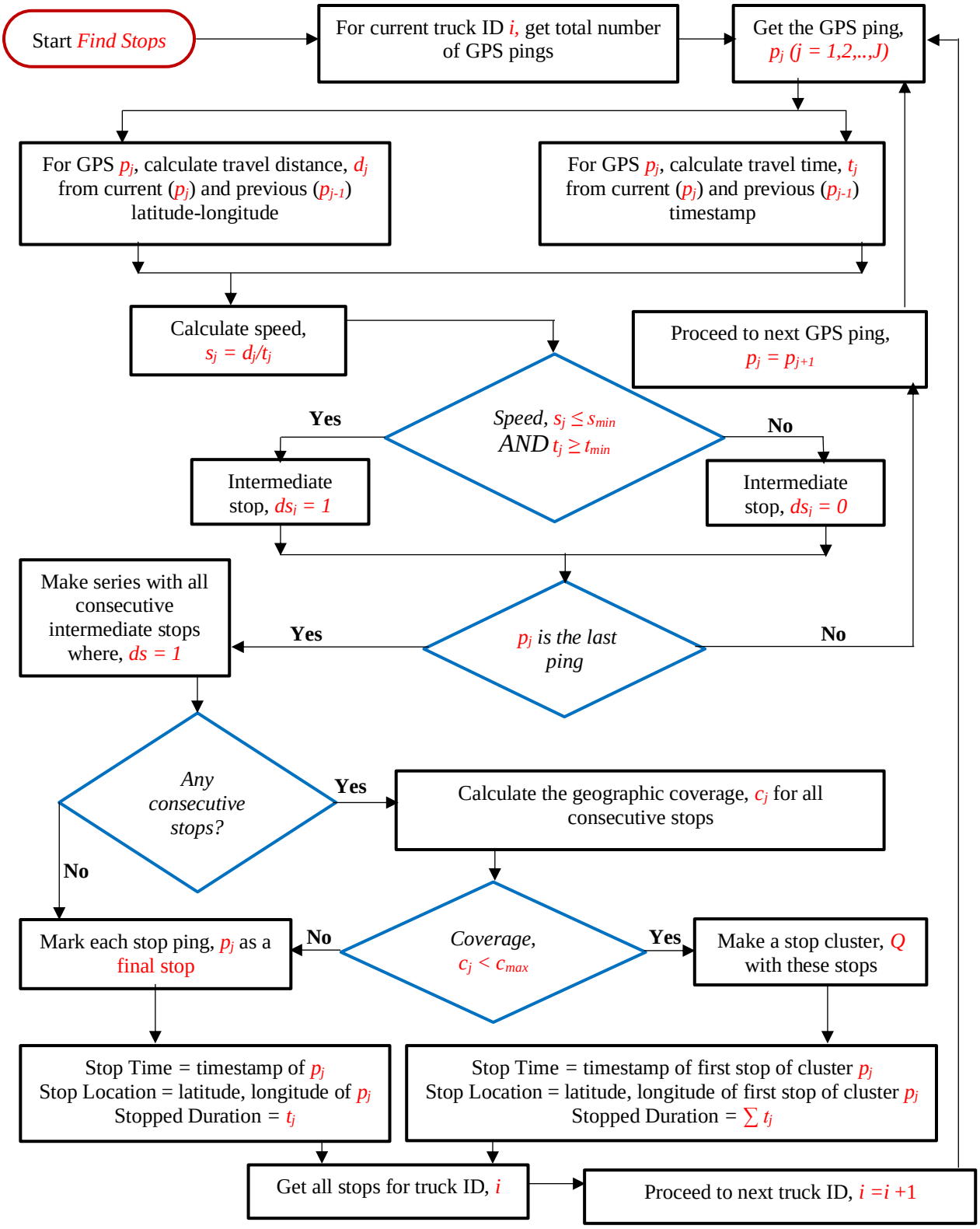


Figure 1.3 Algorithm for *stop identification*

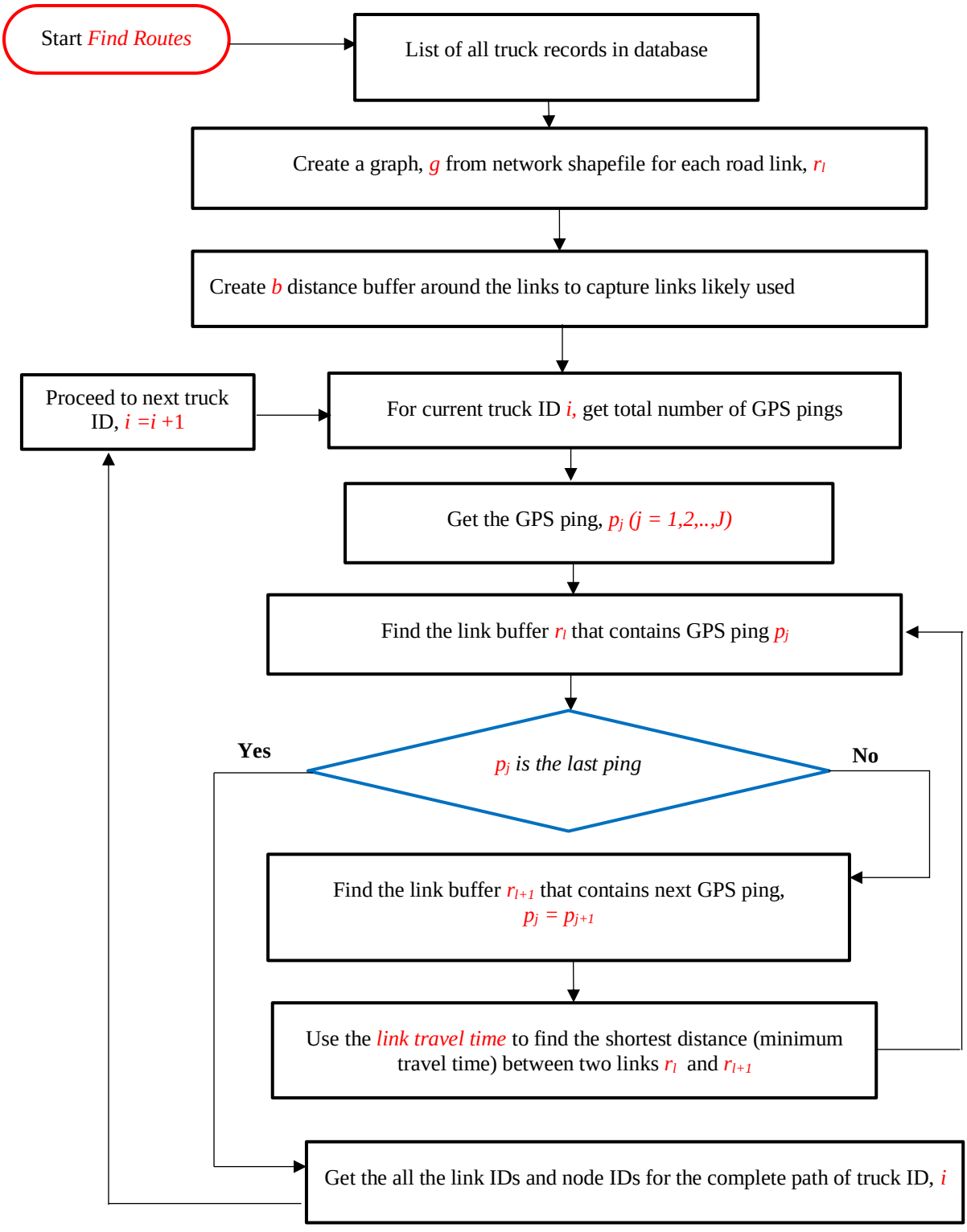


Figure 1.4 Algorithm for *path* identification

1.4.3 Derivation of truck operational characteristics

The stop identification algorithm identified sequential stops and defined stops based on time and duration. The path identification algorithm reconstructed a path as a set of fully connected links defined by link identification number and timestamp. In order to derive operational characteristics, an algorithm was developed to merge results of the stop identification and path identification (Figure 1.5).

First, a serial number, s_j was created for each stop of a truck based on the stop timestamp, t_j (i.e., time and date). Next, each pair of consecutive stops (s_j and s_{j+1}) were classified as a trip, m_j that started with stop s_j and ended with stop s_{j+1}. Thus, each trip was enveloped by two stops, i.e., origin and destination. Stop information (i.e., stop time of day, stop duration, and stop location) of the origin stop were added to each trip.

However, some trips were not bounded by stops. This occurs when a portion of the trip or a stop is outside the boundary of the data sample. For example, for the sample used in this study, all pings inside the Arkansas state boundary plus a ten-mile buffer were used. If a truck had a stop outside the state plus a ten-mile buffer, then we would not be able to observe that stop in our data sample. Likewise, we are unable to observe the remainder of a trip past the state border plus a ten-mile buffer. These "open-ended" trips were still considered by bounding the trip by the state boundary, e.g., the trip is defined from stop location to the state border and vice versa.

Second, path information (i.e., travel length, travel time, speed, and road link characteristics) was combined with stop information for each truck (example in Table 1.1). To combine path and stop data for each truck, the timestamp (t_k) associated with usage of road (r_k) was compared to the stop timestamps (t_j) for trip (m_j) such that if t_k is greater than t_j and smaller

than t_{j+1}. Later, we calculated trip length and trip duration from the combined table (Eq. 1.7 and 1.8).

Table 1.1 Example Results of *Trip Identification* Algorithm

Trip ID	Stop Pair	Stop Time of Day (TOD)	Stop Duration	Stop Location	Road ID	Road Length	Travel Time	Travel Speed	Road Functional Class
					r_1	l_{r1}	t_{r2}	s_{r2}	*Interstate*
m_1	$\{s_1, s_2\}$	tod_{s1}	d_{s1}	l_{s1}	r_2	l_{r2}	t_{r3}	s_{r3}	*Interstate*
					r_3	l_{r3}	t_{r4}	s_{r4}	*Interstate*
m_2	$\{s_2, s_3\}$	tod_{s2}	d_{s2}	l_{s2}	r_4	l_{r4}	t_{r5}	s_{r5}	*Highway*
m_3	$\{s_3, s_4\}$	tod_{s3}	d_{s3}	l_{s3}	r_5	l_{r5}	t_{r6}	s_{r6}	*Highway*
					r_6	l_{r6}	t_{r1}	s_{r1}	*Local*

$$T_{m_j} = \sum_{k=1}^{n} t_{r_k} \tag{1.7}$$

$$L_{m_j} = \sum_{k=1}^{n} l_{r_k} \tag{1.8}$$

Where,

T_{m_j} = Trip duration for trip m_j

t_{r_k} = Travel time for crossing a road link r_k

n = Number of road links in trip m_j

L_{m_j} = Trip length for trip m_j

l_{r_k} = Length of road link r_k

By merging the stop and path identification results we are able to observe trip chains, and thus to derive freight operational characteristics. Based on a review of the literature and the available data, we defined eleven operational characteristics which can be aggregated into four groups (Table 1.2). First, we categorized stops based on stop duration into three categories: less than 30 minutes, 30 minutes to 8 hours, and more than 8 hours. These ranges coincide with

Hours of Service (HOS) regulations for required rest breaks (FMCSA, 2017). For trip length and duration, we categorized trips based on general breakpoints found in the literature defining long and short haul trips. We also considered the Time of Day (TOD) and total number of daily stops as important operational characteristics.

Table 1.2 Operational Characteristics by Group and Type

Feature Group	Features	Variable Type
Stop Duration	1. Number of stops less than 30 minutes 2. 30 minutes to 8 hours 3. More than 8 hours	Discrete
Trip Length	4. Number of trips less than 30 miles 5. 30 miles to 100 miles 6. More than 100 miles	Discrete
Trip Duration	7. Number of trips less than 1 hour 8. 1 hour to 4 hours 9. More than 4 hours	Discrete
Time of Day (TOD)	10. Proportion of daytime stops (6 AM to 6 PM) to all stops 11. Proportion of nighttime stops (12 AM to 6 AM and 6 PM to 12 AM) to all stops	Continuous
Daily Stop	12. Total number of stops in a day	Discrete

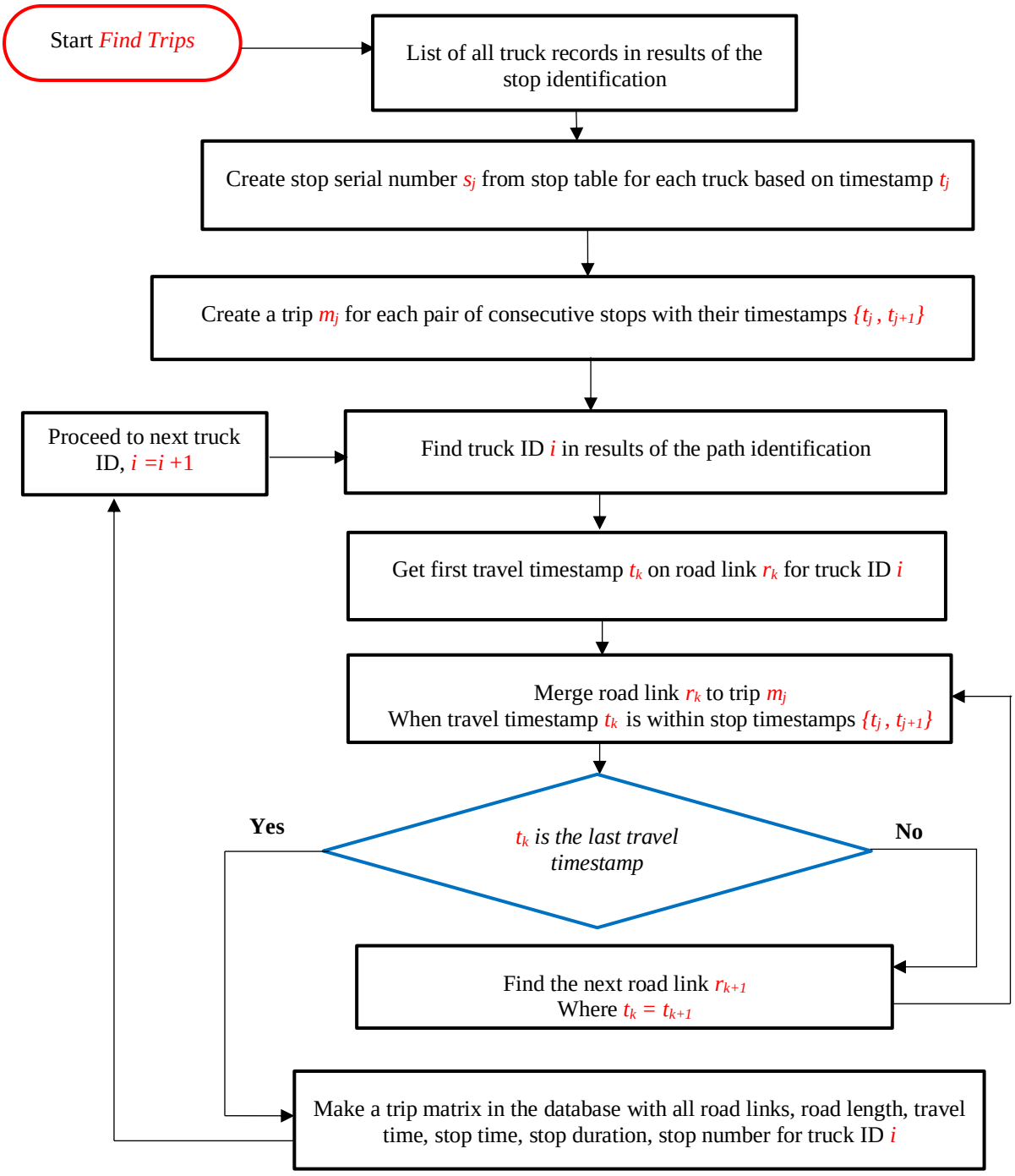


Figure 1.5 Algorithm for *trip identification*

1.4.4 Development of a multinomial logistic (MNL) regression model

A multinomial logistic (MNL) regression model was estimated to define associations between operational characteristics and the probability that a truck was transporting a certain commodity. The premise of the discrete choice model is based on the Random Utility Theory. According to this theory, a decision maker chooses the alternative that yields the highest "utility" (Ben-Akiva and Lerman, 1985; Akar and Clifton, 2009). To extend this theory to prediction of commodity carried, we assume that observed stop and trip characteristics are the result of the commodity being transported. Thus, the probability of a truck transporting commodity *i* can be calculated as:

$$P(i \backslash C_n) = P_r\left(U_{in} \geq U_{jn}\right), \forall j \in C_n \quad (1.9)$$

Where,

U = Utility of the given alternative and

C_n = {*farm products*, *manufacturing*, *mining*, *chemicals*, *miscellaneous mixed*, and *pass-through*}

In our interpretation, the "utility" of alternative *i* can be calculated based on the stop and trip characteristics as:

$$U_{in} = \beta_{in} x_{in} + \varepsilon_{in} \quad (1.10)$$

Where,

U_{in} = Estimated "utility" of alternative (commodity) *i* for driver/truck *n*

x_{in} = Observed stop and trip characteristics

β_{in} = Vector of coefficients of the variables

ε_{in} = Random component, e.g., unobserved or unmeasurable

Under the assumption of the multinomial *logit* model and based on the principle of the utility maximization, the choice probability for alternative *i* can be calculated as:

$$P_n(i) = \frac{e^{Vin}}{\sum_{j \in Cn} e^{Vjn}} \text{ , for all } i \text{ in } j_n \qquad (1.11)$$

Where,

$V_{in} = \beta_{in} x_{in}$

All other terms previously defined.

a. MNL model specification

11 of the 12 operational characteristics derived from the *trip identification* algorithm were used (Table 1.2). To avoid the multicollinearity, the "proportion of nighttime stops" parameter was not included in the model.

Five commodity classes were considered in the model, including:

- *manufactured goods*,
- *farm products*,
- *mining materials*,
- *chemicals*, and
- *miscellaneous mixed*.

Additionally, *pass-through* trucks were considered as a "commodity". This was a necessary addition as *pass-through* trucks represent unique operational behaviors that are not tied to specific commodities. The commodity category *farm products* was chosen as the base category.

b. MNL model estimation

Labeled data is needed to estimate a regression model. In our application, labeled data refers to assigning a commodity carried to each truck trip. To do this, we created a "ground truth" dataset of 2,064 truck trips. The assumption of commodity carried was based on a detailed examination of the truck trip and stops against aerial imageries depicting business and land uses, e.g., Google Satellite images, (Figure 1.6). We were able to distinguish five commodity groups. Commodity groups were treated as the dependent variables and operational characteristics were treated as the independent variables of the MNL model.

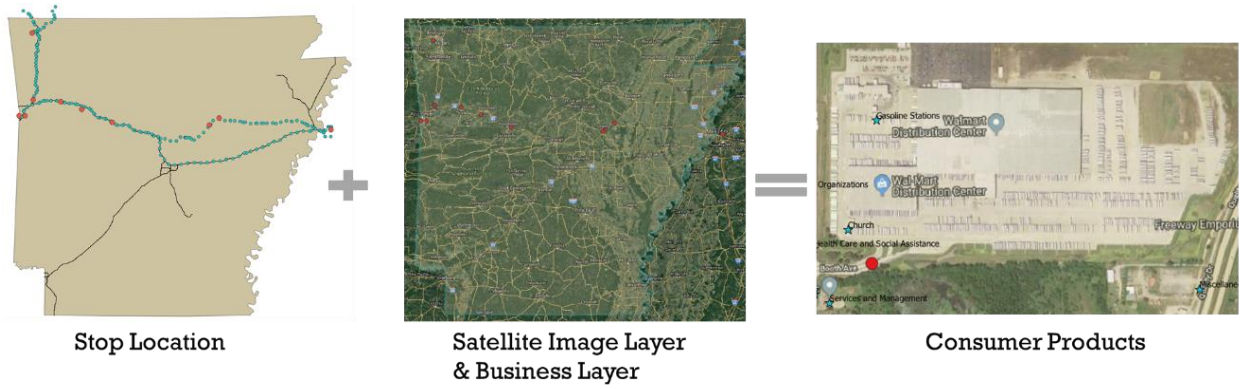


Figure 1.6 Prediction of carried commodity of a truck

Maximum Likelihood Estimation (MLE) was used to estimate the coefficients within the MNL model (Bunch, 1987). At the 95% confidence level, stop duration, trip length, trip duration, stop time of day, and total number of daily stops were found to be significant parameters in predicting commodity carried (Table 1.3).

Table 1.3 Change in Operational Characteristics Based on Commodity Groups

Features			Alternative Commodity Groups (Base: *Farm Products*)				
	Group	Description	*Manuf. Goods*	*Mining Materials*	*Chemicals*	*Misc. Mixed*	*Pass-Through*
Stop Duration	Short break	Less than 30 minutes	2.48***	1.77***	3.64***	4.17***	2.13***
	Pickup/ delivery	30 minutes to 8 hours	2.28***	1.93***	3.33***	3.79***	2.05***
	Long rest break	More than 8 hours	2.43***	2.31***	5.51***	4.03***	1.19
Trip Length	Short-trip length	Less than 30 miles	-0.98***	-0.94***	-4.40***	-1.49***	-1.10**
	Medium-trip length	30 miles to 100 miles	-1.20***	-0.76***	-3.43***	-1.85***	-1.83***
	Long-trip length	More than 100 miles	0.00	0.00	0.00	0.00	0.00
Trip Duration	Short-trip duration	Less than 1 hour	-1.70***	-0.93**	0.82	-2.64***	-1.25
	Medium-trip duration	1 hour to 4 hours	-1.57***	-0.97**	-1.37***	-2.83***	-1.89***
	Long-trip duration	More than 4 hours	0.00	0.00	0.00	0.00	0.00
TOD	Daytime hours	6 AM to 6 PM	0.92***	-0.60***	2.67***	2.18***	-0.18
Daily Stop	Total Stops	Total number of stops in a day	0.03***	0.01***	-0.02*	0.02***	0.04***
Constant			0.32	-0.88***	-6.77***	-4.63***	-1.42***

***significant at 99% confidence level; **significant at 95% confidence level; *significant at 90% confidence level

1.5 Discussion

Knowing the commodity carried by a truck provides insight into its operational characteristics, e.g., number of stops, trip length, time of day travel patterns. Conversely, knowledge of operational characteristics can be used to understand commodity carried by a truck. Because we can derive operational characteristics from GPS data, but cannot observe

commodity carried, we developed heuristic methods to derive operational characteristics from GPS data and then related those characteristics to commodity carried via an MNL model.

According to our MNL estimation, stop time of day, stop duration, trip length, and trip duration were found to be significant operational characteristics predictive of commodity carried. All categories of 'stop duration' were positive and significant for all commodity groups. This indicated that trucks carrying *manufactured products, mining materials, chemicals, misc. mixed*, and those considered *pass-through* had higher number of stops compared to those carrying *farm products*. For instance, if the number of *pickup/delivery* stop increases by one, the log-odds of carrying miscellaneous mixed goods will increase by 4.03 compared to farm products. Alternatively, the log-odds of carrying chemical products will decrease by 4.40 compared to farm products if the number of short length trips increases by one. This denotes that trucks transporting *farm products* had higher number of short length trips compared to those transporting *chemicals*. Additionally, if the number of short duration trips increases by one, the log-odds of carrying farm products will increase by 1.70 compared to manufacturing goods. The model also found that compared to *farm products*, trucks transporting *mining materials* had fewer daytime stops while other commodities had more.

1.6 Conclusion

Although big data like that from GPS is increasingly plentiful, without efficient heuristic methods to extract relevant performance measures we are unable to fully leverage this valuable data source. Methods to derive stop duration, trip length, trip duration, and stop time of day allow us to identify freight activity patterns from big data sources and to link those patterns to commodity carried. While deriving operational characteristics from big data allows us to develop more ubiquitous transportation performance metrics, the link between operational

characteristics and commodity carried serves as critical input for freight demand forecasting (Beagan, Tempesta, & Proussaloglou, 2019).

Our methodology consists of spatial heuristics to identify stop clusters and complete paths of individual trucks from timestamped latitude-longitude points gathered from GPS devices on-board trucks. After deriving stop and path, we can observe trip chains, e.g., sequences of stops and trips. Statistical approaches, namely Multinomial Logit Models (MNL) were employed to determine how operational characteristics like stop time of day and duration, relate to commodity carried. The MNL model identified that stop duration, number of total daily stops, stop time of day, trip length, and trip duration were significant characteristics that could be used to predict commodity carried.

The log likelihood of our MNL model, a general description of the goodness of fit, indicates that there is a room for improvement. This can be attributed to several factors. First, MNL estimation assumes a linear in parameters specification such that operational characteristics should be linearly related to commodity carried. This assumption may not hold true. Advanced machine learning methods such as *K*-means clustering, random forest, and SVM models can better identify patterns, especially non-linear patterns, from large and noisy data like GPS pings (Caruana & Niculescu-Mizil, 2006). Hence, machine learning models are likely more appropriate for this application. Second, MNL specification requires a complete choice set to be specified. We considered only five commodity groups plus a sixth group representing pass through movements. This is not a complete choice set and future work should expand the set of commodities.

The results of this paper can guide public sector engineers and planners to achieve the Transportation Performance Measurement (TPM) goal setting initiatives and requirements set forth in federal transportation legislation.

1.7 Acknowledgement

The authors thank the Arkansas Department of Transportation (ARDOT) for sponsoring the project that led to this paper.

1.8 Authors Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: T. Akter and S. Hernandez; data gathering and processing: T. Akter and interpretation of results: T. Akter, S. Hernandez, and P. Camargo; draft manuscript preparation: T. Akter. All authors reviewed the results and approved the final version of the manuscript.

1.9 References

Akar, G., & Clifton, K. J. (2009). Influence of Individual Perceptions and Bicycle Infrastructure on Decision to Bike. *Transportation Research Record: Journal of the Transportation Research Board, 2140*(1), 165-172. doi:10.3141/2140-18.

Bassok, A., McCormack, E. D., Outwater, M. L., & Ta, C. (2011). Use of Truck GPS Data for Freight Forecasting. Paper presented at the *90th Annual Meeting of the Transportation Research Board.*

Beagan, D., Tempesta, D., & Proussaloglou, K. (2019). *Quick Response Freight Methods*. Retrieved from https://ops.fhwa.dot.gov/publications/fhwahop19057/fhwahop19057.pdf.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete Choice Analysis*. Cambridge, Mass. [u.a.]: MIT Press. Retrieved from http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=000297758&sequence=000002&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Bunch, D. S. (1987). Maximum Likelihood Estimation of Probabilistic Choice Models. *SIAM Journal on Scientific and Statistical Computing, 8*(1), 56-70. doi:10.1137/0908006.

Camargo, P., Hong, S., & Livshits, V. (2017). Expanding the Uses of Truck GPS Data in Freight Modeling and Planning Activities. *Transportation Research Record, 2646*(1), 68-76. doi:10.3141/2646-08.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. Paper presented at the 161-168. doi:10.1145/1143844.1143865 Retrieved from http://dl.acm.org/citation.cfm?id=1143865.

CPCS. (2018). *NCFRP 49 [Final]: Understanding and Using New Data Sources to Address Urban and Metropolitan Freight Challenges*. Retrieved from http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=3593.

Evans, D. L., Kassinger, T. W., Cooper, K. B. & Kincannon, C. L. (2004). 2002 Vehicle Inventory and Use Survey. Retrieved from https://www.census.gov/library/publications/2002/econ/census/vehicle-inventory-and-use-survey.html.

FMCSA. (2017). Summary of Hours of Service Regulations. Retrieved from https://www.fmcsa.dot.gov/regulations/hours-service/summary-hours-service-regulations.

Giovannini, L. (2011). *A Novel Map-Matching Procedure for Low-Sampling GPS Data with Applications to Traffic Flow Analysis* doi:10.6092/unibo/amsdottorato/3898. Retrieved from https://www.openaire.eu/search/publication?articleId=od_____1754::2e76bee797112fda11280f4851def321.

Greaves, S. P., & Figliozzi, M. A. (2008). Collecting Commercial Vehicle Tour Data with Passive Global Positioning System Technology. *Transportation Research Record: Journal of the Transportation Research Board, 2049*(1), 158-166. doi:10.3141/2049-19.

Jing, P. (2018). *Identifying and Modeling Urban Truck Daily Tour-Chaining Patterns* (Doctoral dissertation, Massachusetts Institute of Technology).

Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Vallabhaneni, L., & Nippani, S. (2014). Development of A Tour-Based Truck Travel Demand Model Using Truck GPS Data. Paper presented at the *93rd Annual Meeting of the Transportation Research Board.*

Liao, C. (2009). *Using Archived Truck GPS Data for Freight Performance Analysis On I-94/I-90 From the Twin Cities to Chicago*. University of Minnesota Center for Transportation Studies. Retrieved from https://conservancy.umn.edu/handle/11299/97668.

Ma, X., McCormack, E. D., & Wang, Y. (2011). Processing Commercial Global Positioning System Data to Develop A Web-Based Truck Performance Measures Program. *Transportation Research Record: Journal of the Transportation Research Board, 2246*(1), 92-100. doi:10.3141/2246-12.

McCormack, E. D., Ma, X., Klocow, C., Currarei, A., & Wright, D. (2010). *Developing A GPS-Based Truck Freight Performance Measures Platform*. Retrieved from https://www.wsdot.wa.gov/research/reports/fullreports/748.1.pdf.

Pline, J. L. (1999). *Traffic Engineering Handbook* (5. ed. ed.). Washington, DC: Inst. of Transportation Engineers. Retrieved from http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=009998668&sequence=000001&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Quddus, M., & Washington, S. (2015). Shortest Path and Vehicle Trajectory Aided Map-Matching for Low Frequency GPS Data. *Transportation Research Part C: Emerging Technologies, 55*, 328-339. doi:https://doi.org/10.1016/j.trc.2015.02.017.

Roorda, M. J., Cavalcante, R., McCabe, S., & Kwan, H. (2010). A Conceptual Framework for Agent-Based Modelling of Logistics Services. *Transportation Research Part E: Logistics and Transportation Review, 46*(1), 18-31. doi:https://doi.org/10.1016/j.tre.2009.06.002.

Sharman, B. W., & Roorda, M. J. (2011). Analysis of Freight Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board, 2246*(1), 83-91. doi:10.3141/2246-11.

Thakur, A., Pinjari, A. R., Zanjani, A. B., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board, 2529*(1), 66-73. doi:10.3141/2529-07.

Yang, X., Sun, Z., Ban, X. J., & Holguín-Veras, J. (2014). Urban Freight Delivery Stop Identification with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2411*(1), 55-61. doi:10.3141/2411-07.

Zanjani, A. B., Pinjari, A. R., Kamali, M., Thakur, A., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2494*(1), 87-96. doi:10.3141/2494-10.

Zhao, W., McCormack, E., Dailey, D. J., & Scharnhorst, E. (2013). Using Truck Probe GPS Data to Identify and Rank Roadway Bottlenecks. *Journal of Transportation Engineering, 139*(1), 1-7. doi:10.1061/(ASCE)TE.1943-5436.0000444.

Chapter 2

2 Representative Truck Activity Patterns from Anonymous Mobile Sensor Data

2.1 Abstract

With new sources of big data, it is increasingly possible to practically implement advanced freight forecasting models including activity-based and truck touring models. Such models improve upon traditional trip-based approaches by capturing freight behaviors sensitive to transportation policy and infrastructure changes. A persistent challenge with the use of big data in this context is the ability to generalize a set of representative behaviors to serve as the basis for model calibration and validation from anonymized data depicting the complex behaviors of the population. To address this challenge, we present a two stage methodology to extract unique and representative freight activity patterns from passively collected truck Global Positioning System (GPS). The first stage involved a heuristic-based approach to derive a set of stop and trip characteristics from large-streams of GPS pings. The second stage employed data mining and machine learning techniques to discern common freight activity patterns from the set of defined features. The resulting activity pattern profiles, defined as chains of activities and their trajectories over time and space, allow us to maintain the anonymity of the trucks included in the GPS dataset while providing high-resolution travel profiles- a necessary condition for most data sharing agreements between public agencies and private data providers. Evaluation of our methodology using a GPS data set covering a state-wide region showed six representative daily activity patterns depicting unique truck operations, i.e., long-haul movements with single stop, short-haul home-based movements with multiple stops, and medium-haul home-based movement with one/multiple stops. These activity patterns serve as the critical, and currently missing, data needed to calibrate and validate advanced freight forecasting models. With more advanced

forecasting models reflective of observed freight behaviors, we will be able to more accurately evaluate a wider spectrum of policy and infrastructure scenarios.

2.2 Introduction

Although a number of theoretical Activity Based Models (ABMs) and truck touring models have been developed from as early as 1979 (Adler & Ben-Akiva, 1979), practical implementations have been hindered in part by the unavailability of the data necessary to construct these advanced freight demand forecasting models. In more recent history, growing availability and access to big data from cell phones, Global Positioning Systems (GPS), etc., seemingly closes this data gap. However, we still lack the ability to generalize a set of representative travel patterns from the more complex behaviors of the truck population contained in big data. A representative set of travel patterns is necessary for practical calibration and validation of advanced freight travel demand models. Our study develops a methodology to extract unique, representative, and anonymous truck activity patterns from historical truck GPS data, a common source of big data for freight. In this way, we seek to fill a critical research gap concerning the use of big data for advanced freight demand forecasting.

An activity pattern is defined by start/end times, activity duration, travel duration and length, and sequence of those components. Activity patterns have traditionally been derived from travel surveys (Nepal, Farnsworth, & Pearson, 2005; Nepal, Farnsworth, & Pearson, 2006; Ruan, Lin, & Kawamura, 2012; Allahviranloo, Regue, & Recker, 2017) and, less commonly, from mobile sensors (Chung & Shalaby, 2005). Travel surveys have the benefit of linking activity patterns to demographic characteristics but are limited by smaller sample sizes and temporal scopes, e.g., daily or weekly trip diary formats. It can be difficult to extrapolate activity patterns from a one-day travel survey to the population given the complex decision-making processes

related to trip chaining. Moreover, travel diaries for freight trucks are almost non-existent. For example, the Vehicle Inventory and Use Survey (VIUS) carried out by the FHWA gathered data from fleet managers on annual trip and vehicle characteristics but did not at all resemble a typical trip diary that was needed to recreate travel patterns (FHWA, 2001).

Counter to travel surveys, passively collected mobile sensor data for freight captures a much larger proportion of the truck population and provides continuous spatial and temporal coverage. This data is increasingly available due to the prevalence of on-board or cellphone-based GPS units and, recently mandated, Electronic Logging Devices (ELD). Since mobile sensor data typically represents a large but sampled portion of the population, it has been commonly used as a source of probe vehicle data to measure speeds and travel times. Considering this data depicts high resolution vehicle movements, sometimes on the order of minute to minute position updates, and is potentially available for all trucks, there is a significant power in leveraging it to gain insights into freight activity patterns. A persistent challenge with the use of big data in this context is the ability to generalize a set of representative behaviors to serve as the basis for model calibration and validation from anonymized data depicting the complex behaviors of the population.

To address this challenge, we present a two stage methodology to extract unique and representative freight activity patterns from passively collected truck Global Positioning System (GPS). The first stage involved a heuristic-based approach to derive a set of stop and trip characteristics from large-streams of GPS pings. The second stage employed unsupervised machine learning techniques, namely K-means clustering, to discern common freight activity patterns from the set of defined features. The premise of this study follows from the work of Allahviranloo, Regue, and Recker (2017) for passenger activity travel pattern generation.

Allahviranloo, Regue, and Recker (2017) demonstrated, using survey data, that a limited set of representative daily activity patterns can be extracted from those of the larger population and used for ABM calibration and validation. Our work not only extends this approach to freight activity pattern recognition but leverages anonymous mobile sensor data in place of traditional travel surveys.

2.3 Background

Trucking is and will continue to be the dominate mode of transport for freight in the US with trucks accounting for 64% and 69% of the market by both weight and value, respectively (FHWA, 2018). The Freight Analysis Framework (FAF), the Federal Highway Administration's (FHWA) nationwide freight forecasting model estimates that the weight of freight shipments moved by truck will grow 45% between 2012 and 2045 (FHWA, 2018). Ensuring efficient freight movement through the provision of adequate infrastructure and effective transportation policy is critical for the economy and the environment. To construct, maintain, and operate a transportation system that supports the efficient movement of freight, it is necessary for public transportation agencies accurately model and predict freight travel demands.

A variety of travel demand models, i.e., traditional trip-based, activity-based, and truck touring models, are used to predict freight flows and, in turn, direct effective freight-oriented infrastructure and policy programs. However, the choice of an appropriate model depends on data availability, time and resource allotments, and the need to assess certain infrastructure and/or policy scenarios. Advanced freight forecasting models are increasingly used to predict travel demands as they consider robust behavioral characteristics, operational decisions, and interactions. Advance models, compared to their traditional trip-based predecessors, allow agencies to evaluate a wider variety of infrastructure and policy decisions by incorporating

behavioral models. Activity Based Models (ABMs), for example, forecast travel demand by depicting trip chains of individual agents participating in a set of activities. For freight, activities include initiating/receiving shipments and transporting goods from origin to destination by various modes. Agents may be shippers, receivers, or drivers. The premise of such models, unlike trip-based models, is that travel is derived from the demand to pursue activities. Thus, models that consider trip linkages have the potential to more accurately forecast travel demands by focusing on activity patterns rather than individual trips.

With new sources of big data providing insights into freight travel patterns, it is becoming increasingly possible to practically implement advanced freight forecasting models including activity-based and truck touring models. Key to successfully leveraging big data for advanced travel demand modeling is the ability to (1) derive operational characteristics, (2) extract common activity patterns, and (3) link activity patterns to the population.

2.3.1 Deriving operational characteristics

In order to distill common activity patterns from big data sources like GPS, it is first necessary to extract operational characteristics that define activity patterns. Examples of operational characteristics include trip length, number of trips, speed, travel time, destination, stop location, and stop duration (Zanjani et al., 2015; Liao, 2009). Heuristic approaches for identifying stops ('Stop Identification') and trips('Map Matching') have been developed to derive operational characteristics from large-streams of GPS data (Giovannini, 2011; Thakur et al., 2015; Quddus & Washington, 2015; Camargo, Hong, & Livshits, 2017). Stop-identification refers to finding clusters of pings that relate to a single stop. Available algorithms (Thakur et al., 2015; Camargo, Hong, & Livshits, 2017) used geographic bounding boxes and rule-based approaches to define stop clusters. Map-matching refers to the process of identifying the network

links that correspond to each GPS ping (a latitude, longitude, timestamp tuple). Giovannini (2011) developed an algorithm to re-construct routes from low-sample rate GPS data, e.g., around one mile between pings, using a Bayesian approach (Giovannini, 2011). Quddus and Washington (2015) developed a new weight-based shortest path and vehicle trajectory aided *map-matching* algorithm to determine the network link corresponding to each GPS ping based on proximity, among other factors, for a sparse road network. Further extensions of map-matching, such as that by Camargo, Hong, and Livshits (2017), ensured that the sequence of identified network links constituted a complete path. The *stop identification* and *map-matching* algorithms developed by Camargo, Hong, and Livshits (2017) were used in this paper as they were shown to produce accurate stop locations and routes for GPS data. We applied several modifications to their algorithms to ensure accuracy for denser road networks and less urbanized areas.

2.3.2 Extracting representative activity patterns

Due to the ability to handle complex patterns and noise found in large datasets, machine-learning techniques have been used to extract representative activity patterns from surveys (Allahviranloo, Regue, & Recker, 2017; Jiang, Ferreira, & González, 2012; Allahviranloo & Recker, 2013; Li & Lee, 2017) and mobile sources (Shoval & Isaacson, 2007; YANG, YAO, YUE, & LIU, 2010; Liu et al., 2014). Jiang, Ferreira, and González (2012) applied Principle Component Analysis (PCA) and *K*-means clustering to extract representative groups among weekday and weekend activity patterns from travel surveys. They found eight and seven representative groups for weekdays and weekends, respectively. Allahviranloo and Recker (2013) used Support Vector Machines (SVM) to classify the daily activity patterns of travelers based on trip diary data. Allahviranloo, Regue, and Recker (2017) generated activity patterns from survey data using a combination of affinity propagation and *K*-means clustering. They

defined 12 activity patterns, where the pattern corresponding to long duration work activity was the most prevalent. Also working with survey data, Li and Lee (2017) developed a Probabilistic Context Free Grammar (PCGG) model to analyze and generate daily activity patterns. They found 15 common activity patterns which explained 70% of the behaviors represented by their data sample.

Shoval and Isaacson (2007) used a variety of tracking technologies, i.e., GPS tracking, Cellular Triangulation tracking, assisted GPS tracking, and land-based time difference of arrival (TDOA) tracking, to collect and analyze time-space activity patterns of tourists. They found that GPS devices collected more accurate data than other tracking methods. Like the studies by Allahviranloo and Recker (2013) and Allahviranloo, Regue, and Recker (2017), YANG, YAO, YUE, and LIU (2010) applied SVM methods to determine the individual's travel behavior but used GPS data instead of travel surveys. Features used to train their SVM included activity start time, end time, distance, etc. derived from the GPS data (YANG, YAO, YUE, & LIU, 2010). They were able to distinguish around eight unique activity patterns. Similarly, Liu et al. (2014) used mobile phone data to identify activity types based on travel behavior information, i.e., the timing and frequency of visits to different locations. Liu, Janssens, Cui, Wets, and Cools (2015) developed a model based on profile Hidden Markov Models (pHMMs) to quantify the occurrence probabilities of all the daily activities as well as their sequential order also using mobile sensor data. They found three main patterns dependent on the location of the longest activity duration, i.e., home, work, and non-work clusters, where the non-work cluster had seven sub-clusters. Considering the availability of truck GPS data, there is significant potential in extending the abovementioned techniques to distill activity patterns for freight.

2.3.3 Linking representative activity patterns to the population

To expand representative activity patterns extracted from surveys or samples of mobile sensor data to the population-at-large, it is necessary to link patterns to freight demographic characteristics like industry served and commodity carried. However, commercially available mobile sensor data is typically devoid of demographic data, e.g., anonymized, to protect the privacy and satisfy data sharing agreements between public agencies and private data providers. Jing (2018) attempted to overcome this limitation by concurrently collecting travel diary and GPS data for freight trucks through a tablet-based application. Like traditional travel surveys, this approach was restricted by its smaller sample size (i.e., the survey included only 119 truck drivers in Singapore), bringing into question the ability to extrapolate derived activity patterns to a much larger truck population (Jing, 2018).

Without survey data to provide necessary demographics like trip purpose, commodity carried, or truck type, algorithmic approaches to derive such information from GPS data have been attempted. Kuppam et al. (2014) combined GPS and land use data to derive trip purposes, i.e., goods pickup or delivery, service, return home. They showed that land use at the trip origin was a significant predictor of trip purpose and was able to correlate industry type with trip characteristics like frequency and number of stops. For example, “construction trucks” made fewer stops than “government-related trucks”. Unlike the study by Kuppam et al. (2014) which was able to correlate freight demographics from activity or trip characteristics, Ma, McCormack, and Wang (2011) focused on distinguishing vehicle characteristics from mobile sensor data, which can also be useful for inferring freight demographics. They used GPS data to classify truck trips into access, local, and loop trips based on trip travel distance from the origin to the destination relative to straight-line distances. Similar to these approaches, the methodology

described in this paper connects activity patterns to freight demographics, specifically industry served, by examining land uses at each stop location.

2.4 Methodology

Following a brief discussion of the data requirements, the two major components of the methodology are discussed in this section: (1) derivation of operational characteristics from truck GPS data, and (2) selection, estimation, and validation of unsupervised machine-learning models to discern unique truck activity patterns from operational characteristics.

2.4.1 Data requirements

The methodology described in this paper is suited for large streams of mobile sensor data that contain a unique, but anonymous, vehicle identification number (ID), timestamp, latitude and longitude, point-speed, and heading direction (e.g., azimuth). Pre-processing to remove noise and other inconsistencies in the data are necessary, but not described in this paper as they are dependent on the particular data set used. It is assumed that adequate quality checks will produce 'complete' truck records, defined as those that represent an over the road truck movement with reasonable start and end positions, speeds, and accelerations.

Once cleaned of inconsistencies, GPS data represented as a series of pings should be converted to a series of stops and trips. Heuristic approaches developed by Camargo, Hong, and Livshits (2017) to identify stop clusters and routes from truck GPS data were adapted for this work due to differences in proposed application contexts, i.e., metropolitan area vs statewide region. To define stop locations, rather than identifying the centroid of a cluster of stops (e.g., a group of consecutive pings with minimal speed) as the stop location, we used the first identified ping in the cluster to define each stop location. This ensured that stop locations aligned with physical business locations so that after identifying activity patterns we could assign industry-

served to each pattern. In regard to trip characteristics, modifications to the *map-matching* algorithm by Camargo, Hong, and Livshits (2017) accounted for a dense statewide road network. Use of the All Roads Network of Linear Referenced Data (ARNOLD) (FHWA, 2014) network file in this work, ensures the transferability of results from state-to-state. Because this network was denser than that used by Camargo, Hong, and Livshits (2017), the link buffer distance was altered to improve accuracy in matching GPS pings to network links. Additionally, the modified algorithm defined link cost using estimated free-flow travel time instead of link length. Since ARNOLD does not include speed limits, speed limits were assumed based on functional class. Further details on modifications to the *stop identification* and *map-matching* algorithms can be found in Akter, Hernandez, Diaz, and Ngo (2018).

2.4.2 Operational characteristics as input feature vector

Five operational characteristics were extracted from the GPS data, three relating to stops, i.e., stop time of day, number of stops, and stop duration and two relating to trips, i.e., trip length and trip duration. To derive daily activity patterns, we segmented multi-day travel patterns by day (i.e., from midnight to midnight). For instance, if a unique truck traveled for three days, that truck would be segmented into three independent daily truck records. We adopted this approach to consider situations where a unique truck transported different goods on different days and thus showed different activity patterns.

The daily pattern of each truck was represented by an 11-element feature vector based on operational characteristics (Table 2.1). These features relate to behavioral characteristics assumed to distinguish representative activity patterns. For instance, stops of 'less than 30 minutes' duration captured short-breaks, e.g., food break, restroom, refueling, etc. while stops of '30 minutes to 8 hours' duration captured pickup/delivery stops but not long rest periods (Jing,

2018). Trip length and trip duration were used to identify the types of truck trips. Trip lengths 'less than 30 miles' and/or trip duration 'less than 1 hour' were assumed to represent short-haul truck movements while trip lengths 'more than 100 miles' and/or trip duration 'more than 4 hours' represented long-haul truck movements.

Table 2.1 Features Defined by Operational Characteristics by Group and Type

Feature Group	Features	Variable Type
Stop Duration	1. Number of stops less than 30 minutes 2. 30 minutes to 8 hours 3. More than 8 hours	Discrete
Trip Length	4. Number of trips less than 30 miles 5. 30 miles to 100 miles 6. More than 100 miles	Discrete
Trip Duration	7. Number of trips less than 1 hour 8. 1 hour to 4 hours 9. More than 4 hours	Discrete
Time of Day (TOD)	10. Proportion of daytime stops (6 AM to 6 PM) to all stops 11. Proportion of nighttime stops (12 AM to 6 AM and 6 PM to 12 AM) to all stops	Continuous

2.4.3 Unsupervised machine learning to derive representative activity patterns

A *K*-means clustering model was applied to identify the representative daily activity patterns of trucks. The assumption was that *K*-means clustering could distill the daily activity patterns of the truck population to a relatively small set of representative patterns, as well as to identify the optimal number and compositions of such patterns should they exist (Allahviranloo, Regue, & Recker, 2017).

Unsupervised learning methods find multi-dimensional groups in data represented by multi-dimensional input vectors (Alpaydin, 2014). Among unsupervised learning models (i.e., Hierarchical, DBSCAN, Gaussian Mixture Model, etc.), *K*-means cluster models are appropriate when input variables are numerical, as is the case for the feature vector representing operational patterns (Bishop, 2016). *K*-means clustering algorithms partition the data into *K* number of

clusters in a multidimensional space such that the sum of the squares of the distances of each data point to its closest cluster centroid μ_k is a minimum (Bishop, 2016) (Eq. 2.1). A two-step iterative procedure is used to find optimal cluster assignments. Iterations correspond to successive optimizations with respect to the binary indicator variables for cluster membership (r_{nk}) and the cluster centroid "location' (μ_k). The first step assumed a random value for μ_k for K number of clusters and minimizes J with respect to r_{nk} (Eq. 2.2). In the second step, J is minimized with respect to μ_k, keeping r_{nk} fixed (Eq. 2.3 and 2.4). The first stage of updating r_{nk} and the second stage of updating μ_k correspond respectively to the E (expectation) and M (maximization) steps of the EM algorithm. This two-stage optimization is repeated until convergence (Bishop, 2016).

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}|| x_n - \mu_k||^2 \tag{2.1}$$

$$r_{nk} = \begin{cases} 1 \; if \; k = \arg min_j || x_n - \mu_j||^2 \\ 0 \; otherwise \end{cases} \tag{2.2}$$

$$2\sum_{n=1}^{N} r_{nk}\,(x_n - \mu_k) = 0 \tag{2.3}$$

$$\mu_k = \frac{\sum_n r_{nk}\, x_n}{\sum_n r_{nk}} \tag{2.4}$$

Where,

$\{x_1, \ldots, x_N\}$ = N observations of a random D-dimensional Euclidean variable x

μ_k = Centers of the clusters, where $k = 1, \ldots, K$

r_{nk} = Binary indicator variables, $\{0, 1\}$ describing which of the K clusters the data point x_n is assigned to, where $k = 1, \ldots, K$

A challenge in applying *K*-means clustering is the need to define the number of clusters when there is no a priori knowledge of appropriate value. Several approaches are suggested in the literature to select *K* including i) by the rule of thumb, ii) 'elbow' method, iii) information criterion approach, iv) an information theoretic approach, v) choosing *K* using the silhouette and vi) cross-validation (Kodinariya & Makwana, 2013). Of these methods, the 'elbow' method is the most commonly used and, in this study, produced a logical *K* value (Ng, 2012). The "elbow" method considers the number of clusters *K* as a function of the total within-cluster sum of squares (WSS). A reasonable number of clusters *K* differences when there is minimal change in the total WSS after adding another cluster.

2.5 Results

Four, two-week periods of anonymous truck GPS data representing each quarter of the year (i.e., February, May, August/September, and November) gathered from the American Transportation Research Institute (ATRI) were used to assess the proposed method. The data from the August/September sample was used for algorithm calibration, i.e., setting the *stop identification* and *map-matching* parameters and determining an appropriate number of clusters, while the remaining datasets were used for assessing temporal transferability. In total, there were approximately 338,304,135 pings within the eight-week sample period and the sample was shown to be a representative sample of the total truck population (Corro, Akter, & Hernandez, 2019).

The K-means clustering model was applied to approximately 300,000 daily truck movement records and produced six distinct clusters ($K = 6$) from the 11-element input feature vector (Table 2.2). The number of clusters (K) was varied from one to 15 clusters and the 'elbow' method was applied to determine a reasonable number of clusters (Figure 2.1). Since the

WSS plateaued beyond six clusters, minimal differences in cluster characteristics were observed when more clusters were added. Alternatively, total WSS increased when the number of clusters decreased below six clusters.

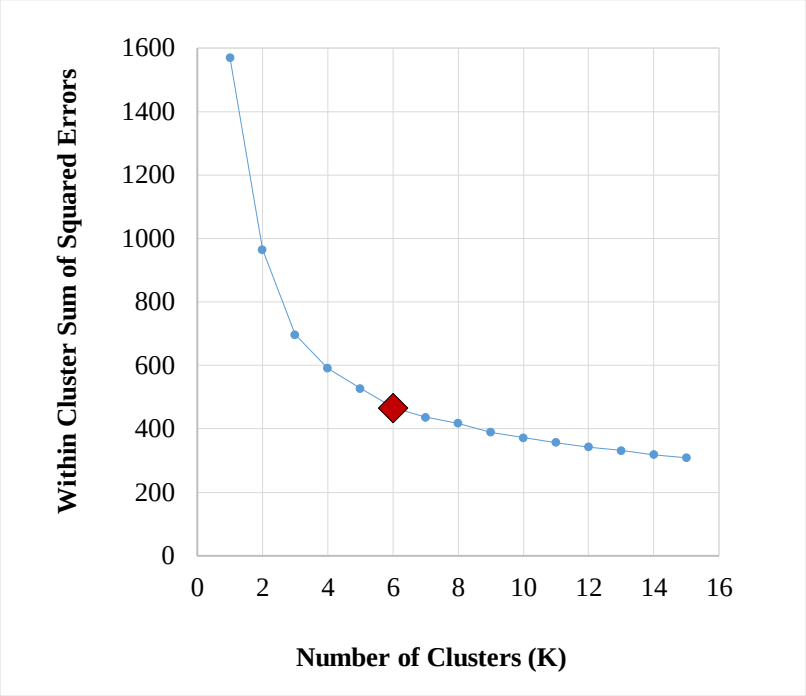


Figure 2.1 Number of clusters based on "elbow method"

The following definitions were adopted to facilitate interpretation of activity patterns represented by each cluster:

- *Short break*: Stop duration less than 30 minutes
- *Pickup/delivery*: Stop duration 30 minutes to 8 hours
- *Long rest break*: Stop duration more than 8 hours
- *Short-trip length*: Trip length less than 30 miles
- *Medium-trip length*: Trip length 30 miles to 100 miles
- *Long-trip length*: Trip length more than 100 miles
- *Short-trip duration*: Trip duration less than 1 hour
- *Medium-trip duration*: Trip duration 1 hour to 4 hours
- *Long-trip duration*: Trip duration more than 4 hours
- *Daytime hours*: 6 AM - 6 PM
- *Nighttime hours*: 12 AM – 6 AM and 6 PM – 12 AM

The highest percentage of sampled trucks (about 32%) were clustered into *Activity Pattern 6* that had one or two daily stops, specifically during daytime hours. Those stops, either a short break or a pickup/delivery, were followed by both short- and long-trip lengths. The second highest percentage (about 20%) of sampled trucks were grouped into *Activity Pattern 4*. Those trucks had one to five daily stops (i.e., short break, pickup/delivery, and long rest break) followed by short-, medium-, and long-trip lengths.

Table 2.2 Centroids of *K*-means Clusters

	Features	Activity Pattern 1	Activity Pattern 2	Activity Pattern 3	Activity Pattern 4	Activity Pattern 5	Activity Pattern 6
Stop duration	1. Less than 30 minutes	2 (7.7)	1 (1.5)	0 (0.6)	1 (2.3)	1 (0.7)	1 (1.4)
	2. 30 minutes to 8 hours	3 (5.9)	1 (2.1)	0 (0.5)	1 (1.3)	1 (0.5)	1 (0.8)
	3. More than 8 hours	1 (0.3)	1 (0.4)	1 (0.0)	1 (0.0)	0 (0.0)	0 (0.0)
Trip length	4. Less than 30 miles	3 (14.9)	1 (3.0)	0 (1.0)	1 (4.3)	0 (0.8)	1 (1.8)
	5. 30 to 100 miles	2 (2.9)	1 (1.1)	0 (0.5)	1 (0.8)	0 (0.3)	0 (0.5)
	6. More than 100 miles	1 (1.1)	1 (0.8)	1 (0.4)	1 (0.5)	1 (0.3)	1 (0.4)
Trip duration	7. Less than 1 hour	4 (15.6)	1 (3.3)	1 (1.2)	1 (4.5)	0 (0.9)	1 (2.0)
	8. 1 to 4 hours	2 (2.2)	1 (1.3)	1 (0.6)	1 (0.9)	1 (0.5)	1 (0.6)
	9. More than 4 hours	0 (0.3)	0 (0.4)	0 (0.3)	0 (0.3)	0 (0.3)	0 (0.2)
TOD	10. Day proportion	0.72 (0.01)	0.45 (0.01)	0.00 (0.00)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)
	11. Night proportion	0.28 (0.01)	0.55 (0.01)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.00 (0.00)
Percentage of trucks within each activity pattern cluster		*9%*	*11%*	*14%*	*20%*	*14%*	*32%*

Note: The variance of the feature within the samples in the cluster is shown in parenthesis.

All stops in *Activity Pattern 4* occurred during daytime hours. Also, we observed that around 14% of sampled trucks were clustered into both *Activity Pattern 3* and *Activity Pattern 5*,

independently. Trucks of *Activity Pattern 3* had long rest breaks during nighttime hours followed by long-trip lengths. Alternatively, trucks of *Activity Pattern 5* had long-trip lengths with no long rest break. Around 11% of sampled trucks in *Activity Pattern 2* had one to four daily stops. Those stops were followed by short- and medium-trip durations. Around 55% of stops in *Activity Pattern 2* occurred during nighttime hours. Further, we found that about 9% of trucks were clustered into *Activity Pattern 1* and had a high number of daily stops (6 to 14 stops in a day). Around 33% of those stops were short breaks and 17% were long rest breaks. Moreover, most of the stops (about 72%) occurred during daytime hours for *Activity Pattern 1*.

2.6 Discussion

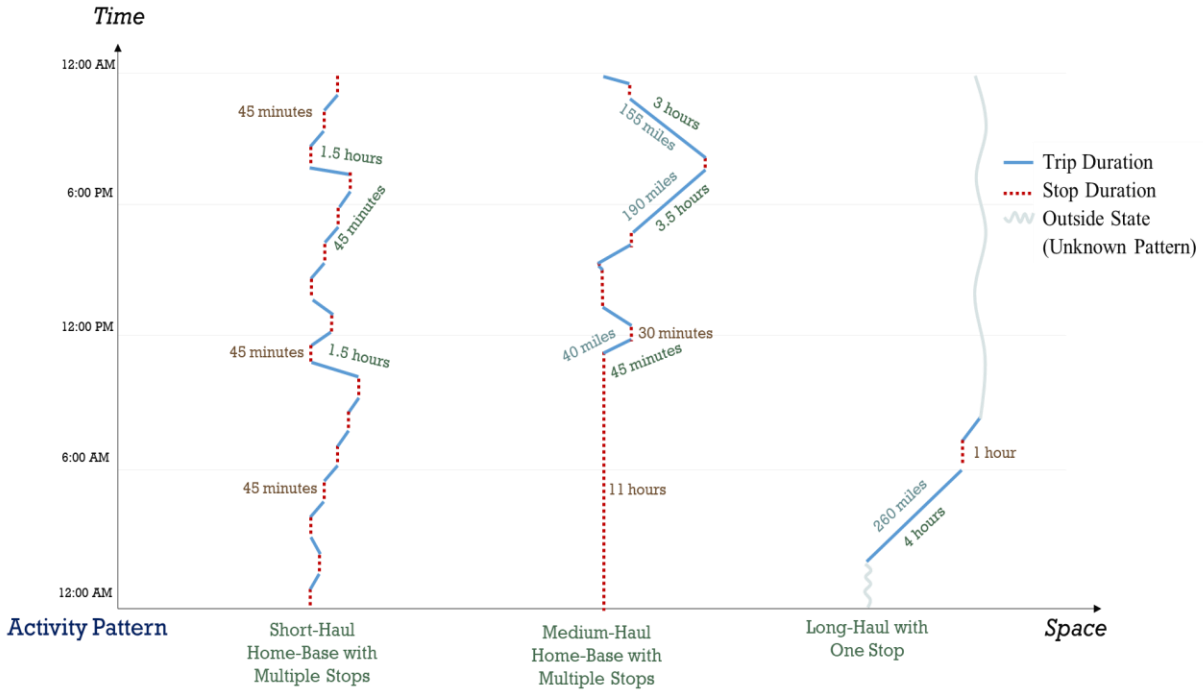
The six representative activity patterns found via K-means clustering using an 11 feature vector depicting operational characteristics can be described according to their spatio-temporal characteristics (Table 2.3). Time-space diagrams depicting changes in location along the horizontal axis (blue lines), duration of activities and travel along the vertical axis (dashed red lines), and portions of the trip that are unknown (grey wavy lines) (Figure 2.2) show distinct patterns. For example, *'Short-Haul Home-Base with Multiple Stops'* (e.g., *Activity Pattern 1*) showed a pattern in which trucks made multiple numbers of stops and returned to their home-base at the end of the day. Trucks labeled *'Medium-Haul Home-Base with One/Multiple Stops'* (e.g., *Activity Pattern 5*) started driving midday after a long rest-break (about 11 hours) followed by a series of short breaks and medium-trip durations (Figure 2.2a). At the end of the day, those trucks also returned to their assumed home base. The last example, labeled '*Long-Haul with One Stop'* (e.g., *Activity Pattern 6*) showed a pattern in which trucks drove through the night and took a short break at 6 AM before resuming their drive across the state (Figure 2.2a). Unlike short and medium-haul movements, these trucks did not return to a home-base by the end of the day. The

grey lines represented unknown portions of the trip. This occurred due to the data sample restriction to truck movements within the state boundary. The remaining activity patterns differed in their number and duration of stops, travel distances, and returns to a home base (Figure 2.2b). As mentioned earlier, *Activity Pattern 2* was similar to *'Short-Haul Home-Base with Multiple Stops'* while *Activity Patterns 3* and *4* were similar to *'Medium-Haul Home-Base with One/Multiple Stops'.*

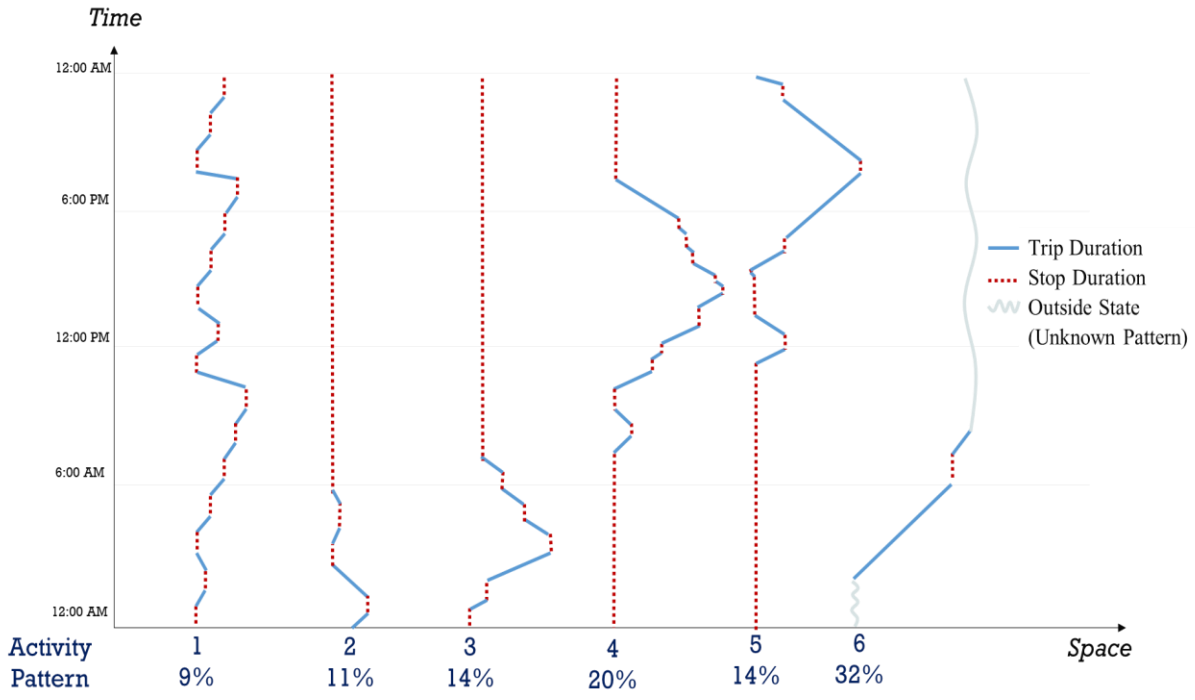
Key to the uniqueness of the six activity patterns in the study was the definition of the feature vector representing the operational characteristics of the trucks. Stop and trip characteristics were two basic operational characteristics that likely varied by commodity carried and industry of the truck. For example, since early morning is the best time to feed hens, trucks carrying chicken feed make multiple short breaks in the morning (before sunrise) followed by short-trip lengths (Waldroup & Hellwig, 2000). Some industries, like mining, operate 24 hours a day and result in a high number of stops and trips throughout the day. By including features that relate to the time of day, stop duration, trip length, and trip duration, we are able to capture these differences in operation that lead to different activity patterns.

Table 2.3 Categorization of Activity Patterns

Activity Pattern	Category Name	Category Description
Activity Pattern 1 *Activity Pattern 2*	Short-Haul Home-Base with Multiple Stops	Trucks have multiple stops followed by multiple short trips and return to home-base within a day
Activity Pattern 3 *Activity Pattern 4* *Activity Pattern 5*	Medium-Haul Home-Base with One/ Multiple Stops	Trucks have one/multiple stops followed by one/multiple medium trips and return to home-base within a day
Activity Pattern 6	Long-Haul with One Stop	Trucks have one (or two) stop followed by one long trip and not return to home-base within a day



(a) Examples of activity pattern types



(b) Activity pattern examples for six clusters

Figure 2.2 Daily activity patterns of freight trucks

A drawback of *K*-means clustering is the a priori need to define the number of clusters. To demonstrate the sensitivity of activity patterns to the selected number of clusters, we examined the activity patterns under assumptions of five (K=5) and seven (K=7) clusters and noted the trends in cluster centroid definitions as we increased the number of clusters beyond seven. With five clusters, *Activity Pattern 5* merged with *Activity Pattern 3*. Thus, we were unable to see subtle differences in medium-haul trips. Specifically, *Activity Pattern 3* had one long-trip duration stop while *Activity Pattern 5* had one short-trip duration followed by a pickup/delivery. Increasing the number of clusters from five to six allowed us to distinguish *Activity Pattern 5* and *Activity Pattern 3*. Increasing from six to seven clusters, on the other hand, divided *Activity Pattern 1* into two clusters. However, the newly created pattern had no meaningful characteristics that would distinguish it as a unique pattern, only a difference in the number of daily stops without changes in the trip length/duration or sequencing among stops. Thus, six clusters were assumed to capture unique and representative activity patterns from the sample.

Variation in the representative activity patterns arose not only due to the selection of the number of clusters but was also found within the samples that comprised each cluster. *Activity Pattern 1*, which represented the lowest percent (about 9%) of daily truck samples, had the highest within-cluster variance for each feature. Other activity patterns had relatively smaller within-cluster variation for each feature. Features with the highest within-cluster variation across all clusters included trips less than 30 miles (feature #4) and trip duration less than 1 hour (feature #7) while the lowest variation was found with stop duration more than 8 hours (feature #3), trips more than 100 miles (feature #6), and trips longer than 4 hours (feature #9). The higher number of short-trips in a day (versus one long-trip) was likely responsible for this variation.

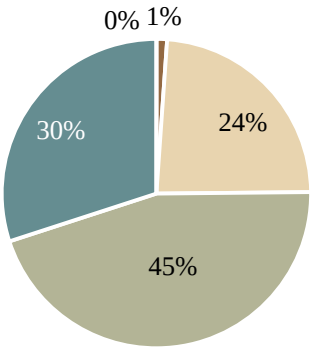
High variation among features in *Activity Pattern 1* explained why increasing the number of clusters leads to further separation of that pattern.

To tie activity patterns distilled from the GPS data sample to those of the larger population for which demographics are known, it was necessary to link each pattern to freight demographics such as commodity or industry type. To create this linkage, 2,064 daily activity patterns were mapped using Google Earth, and the business types of each stop location were examined to determine the industry served by the truck. Industry types were aggregated into five groups defined as follows:

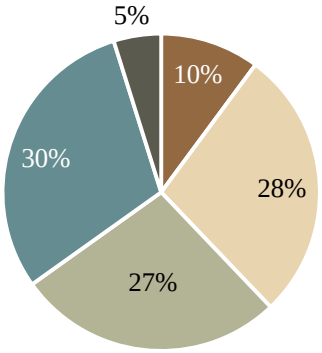
1. *Agriculture* including agriculture and livestock
2. *Materials* including mining, coal, oil/gas, and non-metallic minerals
3. *Consumer products* including food, lumber, and other manufactured products
4. *Equipment and chemicals* including paper, chemicals, concrete, and metals
5. *Pass-through* which included stops at rest areas and gas stations

Each activity pattern cluster consisted of trucks serving multiple industries, however, there was a dominant industry group for several of the activity patterns (Figure 2.3). Of all trucks included in *Activity Pattern 1*, 45% served the materials industry and 30% served the agriculture industry (Figure 2.3a). We assumed this was in line with operations of trucks traveling to and from oil and gas wells to support fracking activity, e.g., many short duration stops and trips with a return to a home base at the end of the day. Further supporting this assumption was the location of stops for *Activity Pattern 1* (i.e., *Short-Haul Home-Base with Multiple Stops*) which align with known oil and gas wells (Figure 2.4a). Those same locations also had businesses related to poultry which tend to generate short-haul truck trips between feed mills, chicken houses, and processing facilities. *Activity Patterns 2 and 3* shared similar distributions among industry types

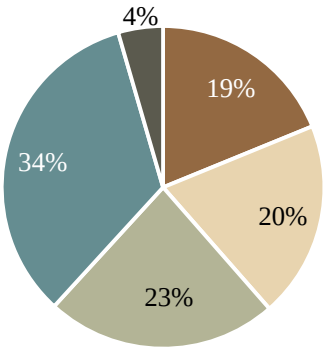
with agriculture representing approximately 30 and 34%, followed by materials representing approximately 27 and 23%, respectively (Figure 2.3b and 2.3c). *Activity Patterns 3* was distinguished by several medium-trip lengths and short breaks with a return to a home base. We assumed the activity patterns related to agriculture in this case were capturing grain production and processing where movements were within the state (i.e., *Medium-Haul Home-Base with One/Multiple Stops*) to and from farms and centralized grain elevators. This was also seen in the relatively heavier volumes of *Activity Pattern 3* trucks in the northeast and northwest regions of the state where farms are located (Figure 2.4b). For materials, we assumed the medium-haul, home based activities captured movements of petroleum between fueling stations. Further, about 55% of trucks following *Activity Patterns 6* represented pass-through movements (Figure 2.3f). The heatmaps of *Activity Pattern 6* (i.e., *Long-Haul with One Stop*) also showed that these trucks had a high concentration of stops in the center region of the state (Figure 2.4c). We considered this pattern as pass-through truck movements that took short-breaks followed by long-trip lengths. The approach of linking activity pattern to industry type is transferable to any geographic extent, although industry types may differ based on the area.



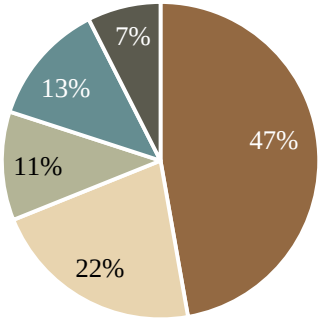
(a) *Activity Pattern 1*



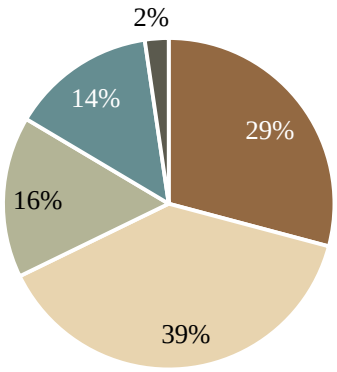
(b) *Activity Pattern 2*



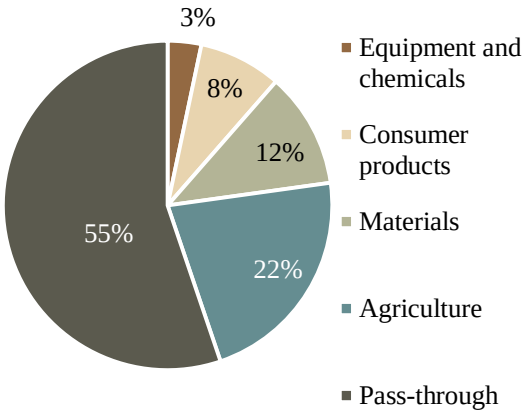
(c) *Activity Pattern 3*



(d) *Activity Pattern 4*

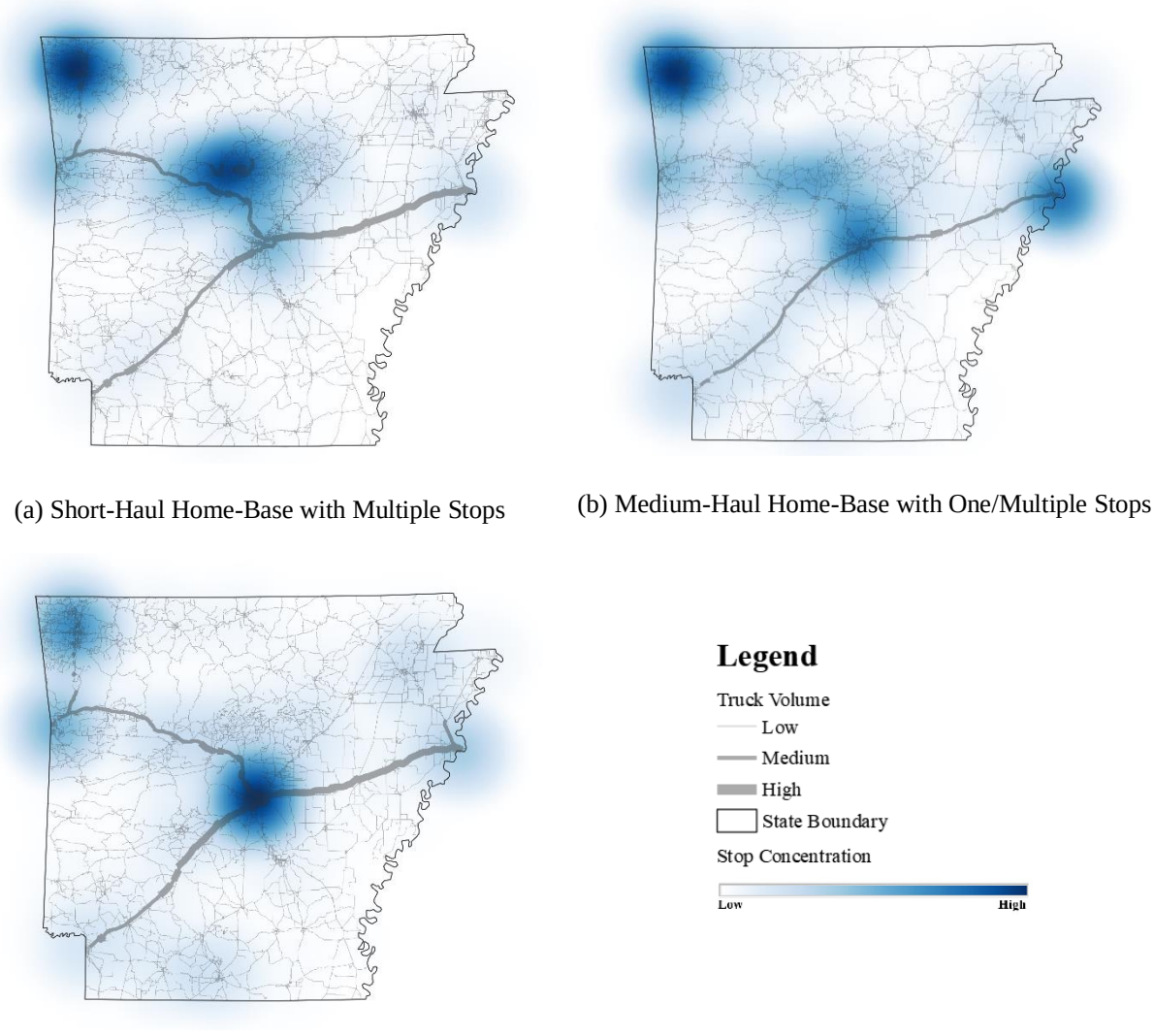


(e) *Activity Pattern 5*



(f) *Activity Pattern 6*

Figure 2.3 Industry types contained in each activity pattern cluster



(a) Short-Haul Home-Base with Multiple Stops

(b) Medium-Haul Home-Base with One/Multiple Stops

(c) Long-Haul with One Stop

Figure 2.4 Stop location concentration by activity pattern

2.7 Conclusions

Transportation agencies rely on freight demand forecasting models to develop, prioritize, and assess future infrastructure and policy scenarios. Advanced freight forecasting models that incorporate behavioral dimensions, including activity-based and truck touring models, allow for a wider range of policy evaluation and more detailed infrastructure planning. To date, such models have been hindered by a lack of relevant and available data. Fortunately, with new

sources of big data evolving in the freight context, it is increasingly possible to practically implement advanced freight forecasting models. Unfortunately, the ubiquity of big data in and of itself does not close this critical data gap. This paper addresses the challenge of using big data for advanced freight travel demand modeling by developing and evaluating a method to extract representative and unique activity patterns from a common source of big data for trucks, e.g. passively collected GPS data.

A two-stage methodology is developed in which daily trip and stop characteristics are extracted from large streams of GPS pings (e.g., latitude, longitude, timestamp) and then used to find common but unique activity patterns defined as series of trips and stops. Heuristic based approaches to determine stop and trip characteristics were used in the first stage that fed into a *K*-means unsupervised clustering algorithm in the second stage. Using a statewide sample of GPS data for evaluation, we identified six activity patterns among 300,000 daily truck records. In relation to advanced freight models like ABMs, by reducing 300,000 daily truck activity patterns to a representative set of six, we aim to enable more efficient model calibration and validation.

About 32% of all trucks included in our statewide GPS sample belonged to the activity pattern cluster representing long-haul movements with a single stop, indicative of pass-through operations. The second most common patterns, approximately 50% in total if combined, captured medium-haul trips with several stops and a daily return to a home base but differed by the time of day in stop and trips took place. The least common pattern depicted short-haul trips with many stops connected by short trips, characteristics of local delivers or local mining operations.

Since truck GPS data used in our study was anonymous, it was not possible to directly "observe" the demographic characteristics (e.g., industry-served or commodity carried) of the

trucks within each representative pattern. Therefore, truck demographic characteristics associated with each activity pattern were inferred through visual comparisons of GPS trajectories and business and land use data. Representative activity patterns linked to industries can improve the ways in which the study extrapolates patterns derived from a sample to the population- a necessary step toward creating the data necessary for advanced freight forecasting models.

In future work, supervised machine learning can be used to predict commodity from operational features such as those described in this paper. For example, through supervised learning techniques, a predictive model can be trained to recognize the operational characteristics (e.g., daily activity patterns) that correspond to particular industries, given a large-enough sample of industry-labeled daily activity patterns. Further, while this study used only truck GPS data to distinguish activity patterns, addition of spatial data depicting business locations and/or land uses and the advent of spatial fusion approaches would allow us to identify the industry associated with each stop and relate it back to commodity specific activity patterns. Ultimately, the developed model demonstrated that activity trajectories for a truck population can be approximated by a small set of representative patterns, containing some core trajectories, and that there are possible correlations among the demographics of commodities and the operational characteristics.

2.8 Acknowledgment

The authors thank the Arkansas Department of Transportation (ARDOT) for sponsoring the project that led to this paper.

2.9 Authors Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: T. Akter and S. Hernandez; data gathering and processing: T. Akter; analysis and interpretation of results: T. Akter and S. Hernandez; draft manuscript preparation: T. Akter. All authors reviewed the results and approved the final version of the manuscript.

2.10 References

Adler, T., & Ben-Akiva, M. (1979). A Theoretical and Empirical Model of Trip Chaining Behavior. *Transportation Research Part B, 13*(3), 243-257. doi:10.1016/0191-2615(79)90016-X.

Akter, T., Hernandez, S., Diaz, K. C., & Ngo, C. (2018). Leveraging Open-Source GIS Tools to Determine Freight Activity Patterns from Anonymous GPS Data. Paper presented at the *2018 AASHTO GIS for Transportation Symposium.*

Allahviranloo, M., & Recker, W. (2013). *Daily Activity Pattern Recognition by Using Support Vector Machines with Multiple Classes* doi:https://doi.org/10.1016/j.trb.2013.09.008.

Allahviranloo, M., Regue, R., & Recker, W. (2017). Modeling the Activity Profiles of A Population. *Transportmetrica B: Transport Dynamics, 5*(4), 426-449. doi:10.1080/21680566.2016.1241960.

Alpaydın, E. (2014). *Introduction to Machine Learning* (3. ed. ed.). Cambridge, Mass. [u.a.]: MIT Press. Retrieved from http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=027423356&sequence=000005&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning* (softcover reprint of the original 1st edition 2006 ed.). New York, NY: Springer. Retrieved from http://deposit.dnb.de/cgi-bin/dokserv?id=efb4c651772a4ab786a75b8fe8f0f0dd&prov=M&dok_var=1&dok_ext=htm.

Camargo, P., Hong, S., & Livshits, V. (2017). Expanding the Uses of Truck GPS Data in Freight Modeling and Planning Activities. *Transportation Research Record, 2646*(1), 68-76. doi:10.3141/2646-08.

Chung, E., & Shalaby, A. (2005). A Trip Reconstruction Tool for GPS-Based Personal Travel Surveys. *Transportation Planning and Technology, 28*(5), 381-401. doi:10.1080/03081060500322599.

Corro, K. D., Akter, T., & Hernandez, S. (2019). Comparison of Overnight Truck Parking Counts with GPS-Derived Counts for Truck Parking Facility Utilization Analysis. *Transportation Research Record, 2673*(8), 377-387. doi:10.1177/0361198119843851.

FHWA. (2001). *Analysis of the Vehicle Inventory and Use Survey for Trucks with Five Axles or More*. Retrieved from https://www.fhwa.dot.gov/reports/tswstudy/vius97.pdf.

FHWA. (2014). *All Road Network of Linear Referenced Data (ARNOLD) Reference Manual*. Retrieved from https://www.fhwa.dot.gov/policyinformation/hpms/documents/arnold_reference_manual_2014.pdf.

FHWA. (2018). Status of The Nation's Highways, Bridges, And Transit Conditions and Performance: 23rd Edition: Part III: Highway Freight Transportation - Report to Congress. Retrieved from https://ops.fhwa.dot.gov/freight/infrastructure/nfn/rptc/cp23hwyfreight/iii_ch11.htm.

Giovannini, L. (2011). *A Novel Map-Matching Procedure for Low-Sampling GPS Data with Applications to Traffic Flow Analysis* doi:10.6092/unibo/amsdottorato/3898. Retrieved from https://www.openaire.eu/search/publication?articleId=od_____1754::2e76bee797112fda11280f4851def321.

Jiang, S., Ferreira, J., & González, M. (2012). Clustering Daily Patterns of Human Activities in The City. *Data Mining and Knowledge Discovery, 25*(3), 478-510. doi:10.1007/s10618-012-0264-z.

Jing, P. (2018). *Identifying and Modeling Urban Truck Daily Tour-Chaining Patterns* (Doctoral dissertation, Massachusetts Institute of Technology).

Kodinariya, T. M., & Makwana, P. R. (2013). Review on Determining Number of Cluster in K-Means Clustering. *International Journal, 1*(6), 90-95.

Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Vallabhaneni, L., & Nippani, S. (2014). Development of A Tour-Based Truck Travel Demand Model Using Truck GPS Data. Paper presented at the *93rd Annual Meeting of the Transportation Research Board.*

Li, S., & Lee, D. (2017). Learning Daily Activity Patterns with Probabilistic Grammars. *Transportation, 44*(1), 49-68. doi:10.1007/s11116-015-9622-1.

Liao, C. (2009). *Using Archived Truck GPS Data for Freight Performance Analysis On I-94/I-90 From the Twin Cities to Chicago*. University of Minnesota Center for Transportation Studies. Retrieved from https://conservancy.umn.edu/handle/11299/97668.

Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., & Cools, M. (2014). Building A Validation Measure for Activity-Based Transportation Models Based on Mobile Phone Data. *Expert Systems with Applications, 41*(14), 6174-6189. doi:https://doi.org/10.1016/j.eswa.2014.03.054.

Liu, F., Janssens, D., Cui, J., Wets, G., & Cools, M. (2015). Characterizing Activity Sequences Using Profile Hidden Markov Models. *Expert Systems with Applications, 42*(13), 5705-5722. doi:https://doi.org/10.1016/j.eswa.2015.02.057.

Ma, X., McCormack, E. D., & Wang, Y. (2011). Processing Commercial Global Positioning System Data to Develop A Web-Based Truck Performance Measures Program. *Transportation Research Record: Journal of the Transportation Research Board, 2246*(1), 92-100. doi:10.3141/2246-12.

Nepal, S. A., Farnsworth, S. P., & Pearson, D. F. (2006). *San Antonio Area Commercial Vehicle Survey Technical Summary*. A report prepared by the Texas Transportation Institute for the Texas Department of Transportation Travel Survey Program.

Nepal, S., Farnsworth, S., & Pearson, D. (2005). *Amarillo Area Commercial Vehicle Survey Technical Summary*. A report prepared by the Texas Transportation Institute for the Texas Department of Transportation Travel Survey Program.

Ng, A. (2012). Clustering with the k-means algorithm. *Machine Learning.*

Quddus, M., & Washington, S. (2015). Shortest Path and Vehicle Trajectory Aided Map-Matching for Low Frequency GPS Data. *Transportation Research Part C: Emerging Technologies, 55*, 328-339. doi:https://doi.org/10.1016/j.trc.2015.02.017.

Ruan, M., Lin, J. (., & Kawamura, K. (2012). Modeling Urban Commercial Vehicle Daily Tour Chaining. *Transportation Research Part E: Logistics and Transportation Review, 48*(6), 1169-1184. doi:https://doi.org/10.1016/j.tre.2012.06.003.

Shoval, N., & Isaacson, M. (2007). Tracking Tourists in the Digital Age. *Annals of Tourism Research, 34*(1), 141-159. doi:10.1016/j.annals.2006.07.007.

Thakur, A., Pinjari, A. R., Zanjani, A. B., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board, 2529*(1), 66-73. doi:10.3141/2529-07.

Waldroup, P. W., & Hellwig, H. M. (2000). The Potential Value of Morning and Afternoon Feeds For Laying Hens. *Journal of Applied Poultry Research, 9*(1), 98-110.

YANG, Y., YAO, E., YUE, H., & LIU, Y. (2010). Trip Chain's Activity Type Recognition Based on Support Vector Machine. *Journal of Transportation Systems Engineering and Information Technology, 10*(6), 70-75. doi:10.1016/S1570-6672(09)60073-8.

Zanjani, A. B., Pinjari, A. R., Kamali, M., Thakur, A., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2494*(1), 87-96. doi:10.3141/2494-10.

Chapter 3

3 Truck Industry Classification from Anonymous Mobile Sensor Data Using Machine Learning

3.1 Abstract

Freight demand forecasting models are used by federal, state, and local transportation agencies to predict future freight flows in efforts to mitigate freight bottlenecks, environmental impacts, and congestion effects. These models are often based on economic forecasts of industry growth and/or commodity production/consumption rates which are then used to estimate expected freight movement, e.g., truck volumes. Unfortunately, there is a lack of data connecting industry served and commodity carried to freight movements which limits the accuracy and usability of freight demand forecasting models. While the private sector (i.e., fleet owners and operators) collects robust data on freight movements including commodity carried, when shared with the public sector, this data is anonymized to protect privacy. As a result, freight movement data is void of industry, commodity, fleet, and driver information. Thus, there is a critical need to re-identify industry served and commodity carried from anonymous freight movement data in ways that maintain privacy standards.

To address this research gap, we developed a classification model using data mining and machine learning methods to predict industry served by a truck from daily activity patterns extracted from truck movement data. Daily activity patterns include stop and trip sequences mined from anonymized truck Global Positioning System (GPS) data. The Random Forest model predicts five industry groups but does not reveal fleet, driver, company, etc. data, providing necessary insight into the relationship between truck movement and economic forecasts. Industry groups include farm products, mining materials, chemicals, manufacturing

good, and miscellaneous mixed goods. From an extensive, manually "groundtruthed" sample of 2064 industry-labeled truck records, the model achieves 90% prediction accuracy. Ultimately, our model allows large streams of truck movement data to be leveraged for freight travel demand forecasting.

3.2 Introduction

Trucking is the dominant mode of transport for freight in the US, moving 64% and 69% of freight by weight and value, respectively (FHWA, 2018). It is predicted to continue to be the dominant mode according to the Freight Analysis Framework (FAF4), the Federal Highway Administration's nationwide freight forecasting model which estimates that the weight of freight shipments moved by truck will grow 45% between 2012 and 2045 (FHWA, 2018). To accommodate this projected growth, it is critical that transportation agencies identify infrastructure and policy solutions to mitigate forecasted freight bottlenecks, environmental impacts, and congestion effects.

To identify effective infrastructure and policy solutions, transportation agencies often develop long-range freight demand forecasting models to predict freight flows to 20 and 40 year planning horizons (Beagan, Tempesta, & Proussaloglou, 2019). In these models, truck movements are estimated as a product of projections of underlying commodity flows such that truck demand is sensitive to economic forecasts (Chow, Yang, & Regan, 2010). This requires knowledge of the relationship between truck movements and commodity carried.

Moreover, freight demand forecasting models that incorporate behavioral representations of truck activity alongside economic based forecasts enable transportation agencies to assess a wide variety of infrastructure and policy solutions. For example, if a model includes a representation of a driver's sensitivity to road pricing during route selection, it is possible to use

the model to assess various tolling policies. Advanced freight forecasting models with behavioral representations of freight movement were proposed as early as 1979 (Adler & Ben-Akiva, 1979). However, practical implementations have been limited by the inability to link truck movements to underlying industry served or commodity carried.

While the private sector (i.e., fleet owners and operators) collects robust data on freight movements including commodity carried, when shared with the public sector for model development, this data is anonymized to protect privacy. As a result, truck movement data available for freight travel demand model development is void of industry, commodity, fleet, and driver information. Thus, there is a critical need to re-identify industry served and commodity carried from anonymous freight movement data in ways that maintain privacy standards.

To address this research gap, we developed a classification model using data mining and machine learning methods to predict industry served by a truck from daily activity patterns extracted from truck movement data. The model predicts five industry groups but does not reveal fleet, driver, company, etc., providing necessary insight into the relationship between truck movement and economic forecasts without violating privacy. Ultimately, our model allows large streams of truck movement data to be leveraged for freight travel demand forecasting.

Our methodology is divided into two approaches: (a) extracting daily activity patterns of trucks from mobile sensor data, and (b) developing a truck industry classification model based on truck operational characteristics. Daily activity patterns include stop and trip sequences mined from anonymized truck Global Positioning System (GPS) data. Considering activity patterns differ by industry served, activity patterns are used to predict industry served. Input features to the supervised machine learning model, namely a Random Forest Model, depict the stop and trip sequences of a truck. Each stop is associated with a particular industry or land use

derived from spatially merging stop locations (from the GPS data) with business locations (from a variety of publicly available spatial data sets). Our classification model can predict the industry served of a truck based on its operational characteristics without disclosing identifiable information like company name. This provides the missing step needed to use mobile sensor data, like that from GPS, for freight travel demand forecasting models.

The paper is organized as follows. The Background section summarizes the most related previous studies to this study. The Methodology section details the data requirements and model specification. The Results section shows the performance of the developed classification model. The paper concludes by highlighting significant findings, noting limitations, and suggesting future improvements.

3.3 Background

Despite being the key component for developing freight travel demand forecasting models, current and historical data on freight truck movements are extremely limited (Beagan, Tempesta, & Proussaloglou, 2019). Public data sources such as the Federal Highway Administration's (FHWA) Freight Analysis Framework (FAF) or the Bureau of Transportation Statistic's Commodity Flow Survey (CFS) contain predictions and observations of freight and commodity flows at the national level (FHWA, 2019a). However, this data is highly spatially aggregated making it difficult if not useless for state and regional planning. For example, FAF, built from CFS data, divides the US into only 123 zones, of which most states are represented by a single zone (FHWA, 2019b). With such aggregated data, it is a challenge for states to use the model or its data assess policy and infrastructure solutions that take place at the state or regional level like for example a local tolling program.

Since most state, regional, and local level planning agencies cannot make effective use of national data resources like FAF and CFS, these agencies must produce and/or acquire their own comprehensive datasets through local establishment surveys, travel diary surveys, roadside intercept surveys, and vehicle classification counts. However, such data sources can be expensive and as a result are often limited in scope (Beagan, Tempesta, & Proussaloglou, 2019).

While private sector data on fleet operations and vehicle movements can be difficult to obtain due to privacy concerns and confidentiality issues, it is becoming both increasingly available and cost effective for state, regional, and local transportation agencies to access (Beagan, Tempesta, & Proussaloglou, 2019). Mobile sensors like Global Positioning System (GPS), crowd-sourced cell phones, and Electronic Logging Devices (ELD) provide detailed depictions of vehicle movements over space and time. Specifically, GPS devices are capable of identifying time-space activity patterns more accurately than other tracking methods (e.g., cellular triangulation tracking) (Shoval & Isaacson, 2007). Advances in the resolution and availability of big data from on-board GPS devices in trucks represents an opportunity to gather freight movement data at spatial resolutions suitable for state, regional, and local freight travel demand model development. For instance, American Transportation Research Institute (ATRI) has a truck GPS database that contains billions of truck data points from more than 500,000 unique vehicles spanning more than 10 years (ATRI, 2019). However, methodological advances are still needed to extract operational characteristics from large and noisy GPS data and re-identify commodity or industry while protecting privacy agreements.

In this section, we summarize prior efforts to (1) extract operational characteristics from mobile sensor data, (2) understand the link between freight operational characteristics and

activity patterns, and (3) use machine learning techniques for data mining and classification of GPS data.

3.3.1 Operational characteristics from mobile sensor data

GPS data, primarily collected by third parties, usually contains only geographic position and timestamp of vehicles, e.g., pings. Ping data, however, does not provide necessary insight into daily operations. For example, it does not explicitly depict locations of stops made by the truck nor does it include what routes were taken between stops. Hence, extracting operational characteristics like trip length, number of trips, speed, travel time, destination, stop location, and stop duration from GPS data is necessary if it is to be used for freight demand forecasting model development.

Stop-identification and *map-matching* are two popular algorithms that identify stops and trips from large streams of GPS data, respectively (Giovannini, 2011; Kuppam et al., 2014; Thakur et al., 2015; Quddus & Washington, 2015; Camargo, Hong, & Livshits, 2017). *Stop-identification* refers to finding clusters of GPS pings that relate to a single stop while *map-matching* refers to the process of identifying the network links corresponding to each ping. Operational data resulting from *stop-identification* and *map-matching* algorithms applied to GPS data include truck speed, travel time, volume, destination, stop location, and stop duration (Liao, 2009; Ma, McCormack, & Wang, 2011).

Methods to derive operational characteristics from GPS data rely on heuristic approaches that differ in their defined parameters for detecting or grouping stops, for example (Zanjani et al., 2015; Liao, 2009; Ma, McCormack, & Wang, 2011). Kuppman et al. (2014) identified a stop if it had speed less than a threshold (e.g., 5 mph) (Kuppam et al., 2014). Geographic bounding boxes and rule-based approaches were also used to identify stop from GPS data (Thakur et al., 2015;

Camargo, Hong, & Livshits, 2017). Giovannini (2011) developed an algorithm that re-constructed routes from infrequent ping data (~1 mile between each ping). He used a Bayesian approach of maximum likelihood for map-matching (Giovannini, 2011). Similar to Giovannini (2011), Quddus and Washington (2015) developed a map-matching algorithm based on shortest path estimation using low-frequency GPS data that determined the corresponding network link to each GPS ping based on proximity, among other factors, for a sparse road network. Further extensions of map-matching ensured that the sequence of identified network links constituted a complete path (Camargo, Hong, & Livshits, 2017). The algorithms developed by Camargo, Hong, and Livshits (2017), applied to a metropolitan area, identified stop time of day, stop location, stop duration, stop coverage, speed, travel time, road link, and road length. Due to the similarities in frequency of GPS pings and transportation network density, we leveraged the stop-identification and map-matching algorithms of Camargo, Hong, and Livshits (2017) in this study.

3.3.2 Link between freight operational characteristics and activity patterns

Operational characteristics refer to stop and trip characteristics such as number of daily stops, stop duration, stop time, trip length, and trip duration of trucks that relate to industry practices. Activity patterns refer to the sequence of operational characteristics over the course of a day or tour that define different industry practices such as long-haul and short-haul operations, pass-through, local, and loop trips, and pickup/delivery, service, and home-based stops. A large body of research has investigated the link between freight operational characteristics and activity patterns. Using GPS data from commercial vehicles, Ma, McCormack, and Wang (2011) classified truck trips into three categories—access trips, local trips, and loop trips— based on trip travel distance. They observed that an access trip had a distinct origin and destination with a

stop longer than three minutes. Besides, they found that the average travel distance for a local trip was less than 0.5 mile while the average travel distance for a loop trip was at least two times larger than network distance between the origin and destination. Both local trips and loop trips did not have a stop longer than three minutes (Ma, McCormack, & Wang, 2011). In a similar study, Zanjani et al. (2015) identified the types of trucks such as light, medium, and heavy trucks from GPS data based on their trip length and the number of trips. They found that a light-duty truck (e.g., local delivery and distribution) made more than five trips per day and none were more than 100 miles in length.

Algorithmic approaches were also used to derive trip purpose, commodity carried, or truck type from GPS data. Kuppam et al. (2014) used a series of discrete choice models to estimate a tour-based model for industries, e.g., retail, farming, household, and industrial. Their model predicted the purpose of each stop and the location of the next stop for different industries. Pickup or delivery, service, and home-base were three examples of stop purposes. The purpose of a truck tour was assumed from the type of land use and it was found that land use of the truck origin had a significant effect on stop purpose. For instance, if a truck's origin location was trade business, the truck was likely to be a retail truck. They also found that the time of day of a stop depended on the purpose of the previous stop. Similar to these approaches, our methodology examines land uses and business types of each stop location to infer the industry served by the truck.

Previous studies have also used both travel diary and GPS data concurrently for freight trucks to overcome the limitation of GPS data (e.g., anonymity) (Jing, 2018). Like traditional travel surveys, the approach of Jing (2018) was limited by its small sample size; the survey included only 119 truck drivers in a large urban area. The small sample size limits the ability to

extrapolate activity patterns derived from the sample to the much larger truck population (Jing, 2018).

*3.3.3 M*achine learning techniques for data mining and classification of GPS data

Advanced analysis techniques such as data mining and machine learning are adept at handling complex patterns and noise typical to large datasets like GPS (Beagan, Tempesta, & Proussaloglou, 2019). These techniques enhance prediction compared to statistical models by addressing higher dimensional and nonlinear relationships among variables (Mortazavi et al., 2016). Several researchers have used machine learning methods to extract representative activity patterns from surveys (Allahviranloo, Regue, & Recker, 2017; Jiang, Ferreira, & González, 2012; Allahviranloo & Recker, 2013; Li & Lee, 2017) and mobile sources (Shoval & Isaacson, 2007; YANG, YAO, YUE, & LIU, 2010; Liu et al., 2014).

Jiang, Ferreira, and González (2012) found eight representative groups for weekdays and seven for weekends from travel surveys after applying Principle Component Analysis (PCA) and *K*-means clustering. Li and Lee (2017) developed a Probabilistic Context Free Grammar (PCGG) model that found 15 common activity patterns and explained 70% of the behaviors represented by their data sample. Also working with survey data, Allahviranloo and Recker (2013) classified the daily activity patterns of travelers based on trip diary data using Support Vector Machines (SVM) techniques to assist activity-based travel demand models. They used two classification techniques in their study: sequential multinomial logistic regression model (MNL) and sequential support vector machines for multiple classes (*K*-SVM). They showed that *K*-SVM models had higher accuracy than MNL models for discerning activity types of passenger trip chains. In another study, Allahviranloo, Regue, and Recker (2017) applied *K*-means clustering with a

combination of affinity propagation on survey data and found long-duration work activity as the most prevalent activity pattern of 12 defined patterns.

Like the studies by Allahviranloo and Recker (2013) and Allahviranloo, Regue, and Recker (2017), YANG, YAO, YUE, and LIU (2010) applied SVM methods to determine the individual's travel behavior but used GPS data instead of travel surveys. Features used to train their SVM included activity start time, end time, and distance, derived from the GPS data (YANG, YAO, YUE, & LIU, 2010). They were able to distinguish eight unique activity patterns. In another study, Yang, Sun, Ban, and Holguín-Veras (2014) identified freight delivery stop from GPS data using the SVM learning method. They used three parameters: stop duration, the distance to the center of the city, and the binary distance to a stop's closest bottleneck as the input feature of the SVM model and yielded a high accuracy of their model with an average error rate of 0.2%. Similarly, Sharman and Roorda (2011) used GPS data to identify the destinations of freight trucks. They applied partitioning methods and hierarchical agglomerative methods to link GPS data to driver records and developed an agent-based travel demand model for commercial vehicles. Moreover, mobile sensor data were used in studies to identify activity types based on travel behavior information, i.e., the timing and frequency of visits to different locations (Liu et al., 2014). Liu et al. (2014) developed a model based on profile Hidden Markov Models (pHMMs) to quantify the occurrence probabilities of all the daily activities as well as their sequential order. They found three major patterns (i.e., home, work, and non-work clusters) depending on the location of the longest activity duration where the non-work cluster had seven sub-clusters.

Gaussian processes (GPs) and ε-support vector machines (ε -SVMs) were also used to predict truck trips with less computational effort compared to multilayer feedforward neural

network (MLFNN) model (Xie & Huynh, 2010). Xie and Huynh (2010) used two Kernel based supervised machine learning methods (GPs and ε -SVMs) to predict daily truck volume at a seaport terminal. Likewise, Sun and Ban (2013) used SVM with quadratic kernel functions to classify general trucks from passenger cars using GPS data. They used average speed, speed variation, and acceleration features as the input variables of the classifier. They found that the average misclassification rate of their model is about 1.6% and 4.2% for the training data and the testing data, respectively (Sun & Ban, 2013).

Besides SVM, other supervised machine learning techniques (e.g., random forest) were used to predict the purpose of truck stops from GPS data (Sarti et al., 2017). Sarti et al. (2017) used heterogeneous GPS data of commercial fleets from diverse industries and developed a random forest model to classify the purpose of stops for commercial vehicles. They classified two types of stops, i.e., work-related stop and non-work-related stop using three types of input features. Stop characteristics (e.g., stop duration, stop time), point of interest (e.g., bank, university), and stop cluster (e.g., land use type) were three input features of their model.

There is significant potential in extending the above-mentioned techniques to distill activity patterns from large samples of truck GPS data. A number of these research efforts use GPS data to classify freight trucks based on their operational characteristics. Some of the previous studies used machine-learning techniques to classify vehicles based on their stop purposes. However, there is still a need to predict commodity carried and industry served of freight trucks so that GPS data can be used to develop and validate freight travel demand forecasting models.

3.4 Methodology

The methodology consists of three approaches: (1) derivation of freight operational characteristics from GPS data, (2) identification of business locations for freight movements, and (3) development of freight industry class model (Figure 3.1).

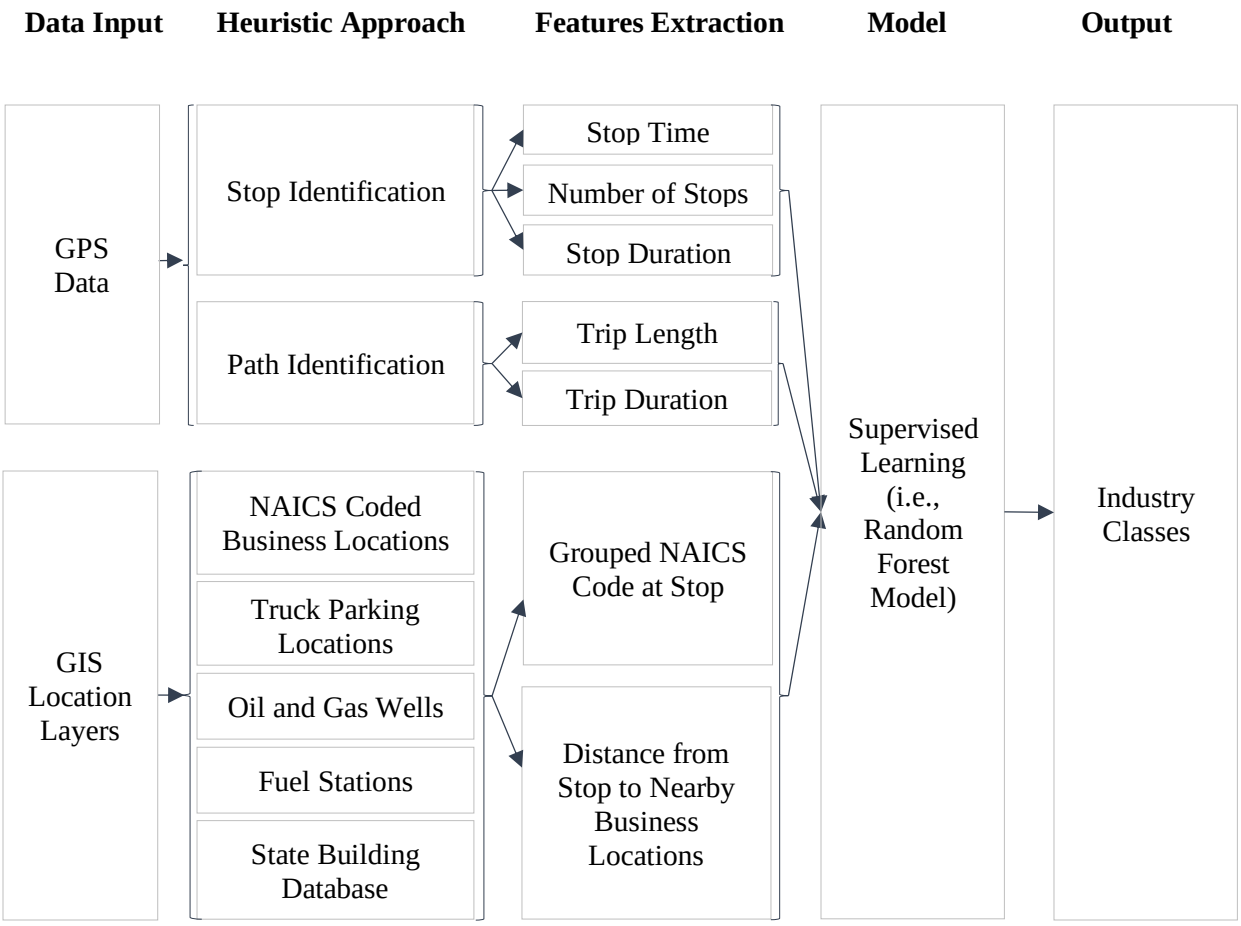


Figure 3.1 Steps to industry classification model

The first and the second approaches make use of spatial heuristics while the third approach employed a supervised machine learning (i.e., random forest model). In this section, we first describe the structure of the data and then explain our heuristic approaches. Next, we introduce techniques adopted to merge the derived operational characteristics with probable

business locations. Finally, we describe our classification model and how we developed a labeled dataset for training and evaluation.

3.4.1 Data requirements and heuristic approaches

Three types of data, (a) mobile sensor data, (b) road network data, and (c) business establishment data were used to conduct this study. Mobile sensor data, i.e., GPS data, contained a unique, but anonymous vehicle identification number (ID), timestamp, latitude and longitude, point-speed, and heading direction (e.g., azimuth). This data required adequate data quality checks to produce "complete" truck records with reasonable start and end positions, speeds, and accelerations. Along with GPS data, we used the All Roads Network of Linear Referenced Data (ARNOLD) network data in this work to ensure the transferability of results from state-to-state (FHWA, 2014).

We used business establishment data from ESRI which contains a comprehensive list of businesses (ESRI, 2019). This data contains name, location, franchise code, industrial classification code, number of employees, and sales of businesses (see example of data in Figure 3.2). Based on the North American Industry Classification System (NAICS) code, we grouped the data into 31 business categories as follows:

1. Agriculture
2. Livestock
3. Forestry
4. Fishing
5. Mining
6. Metal and non-metal
7. Electrical
8. Water and sewerage
9. Building materials
10. Heavy construction
11. Chemicals
12. Plastic and rubber
13. Auto parts and equipment
14. Hospital and medical
15. Miscellaneous consumer products
16. Clothing and accessories
17. Beverage and tobacco
18. Paper
19. Paint
20. Merchandise stores
21. Transportation and warehouse
22. Computer and information
23. Finance and insurance
24. Public administration
25. Waste collection
26. Education
27. Recreation
28. Food stores
29. Service
30. Unclassified
31. All others

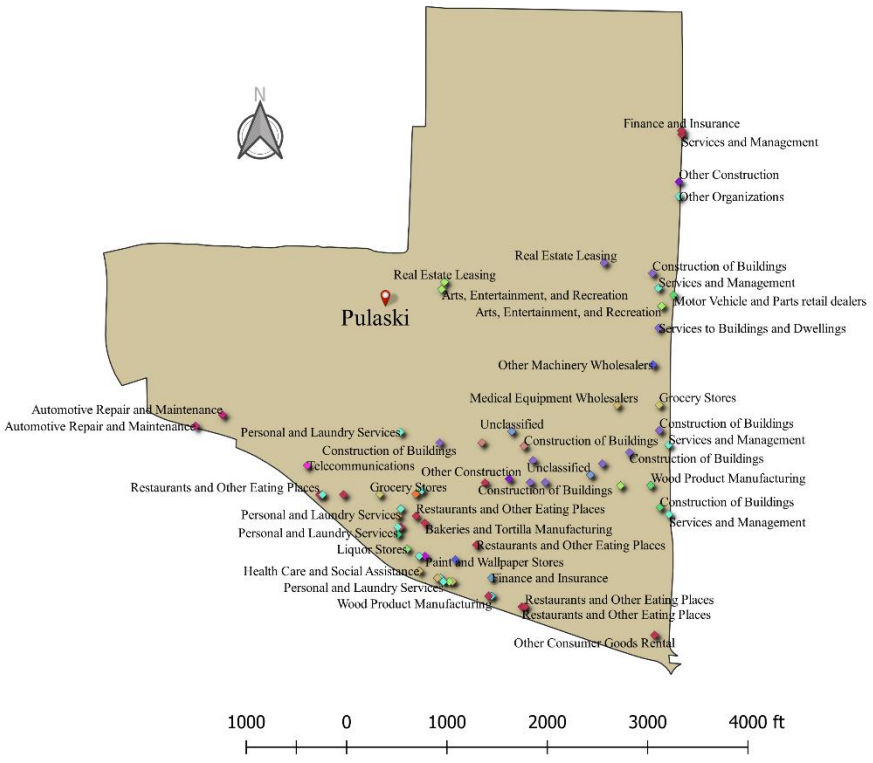


Figure 3.2 Example of business establishments in a Traffic Analysis Zone (TAZ)

After data collection, two heuristics approaches, *stop identification* and *map-matching* algorithms, developed by Camargo, Hong, and Livshits (2017), were applied to derive operational characteristics including number of stops, stop time of the day, stop duration, stop coverage, stop location, trip length, and trip duration. Since we used the algorithms of Camargo, Hong, and Livshits (2017) in a less urbanized statewide region with a denser road network, several modifications to the algorithms were required to ensure accuracy. Further details on modifications to the stop identification and map-matching algorithms can be found in Akter, Hernandez, Diaz, and Ngo (2018).

3.4.2 Operational characteristics and probability matrix of industry class

We extracted freight operational characteristics such as stop duration, stop location, stop coverage, the number of stops, stop time of day, speed, travel time, travel distance, road

functional class (e.g., interstates, highways, and local roads), trip length, and trip duration from the truck GPS data. Five of those operational characteristics (i.e., number of stops, stop time of day, stop duration, trip length, and trip duration) were used as input features for our classification model. Later, we segmented multi-day travel patterns by day (i.e., from midnight to midnight) to capture daily activity patterns of freight trucks. For instance, a unique truck would be segmented into three independent daily truck records if it traveled for three days (Figure 3.3). This approach tackled the situation where a unique truck transported different goods on different days and showed different activity patterns.

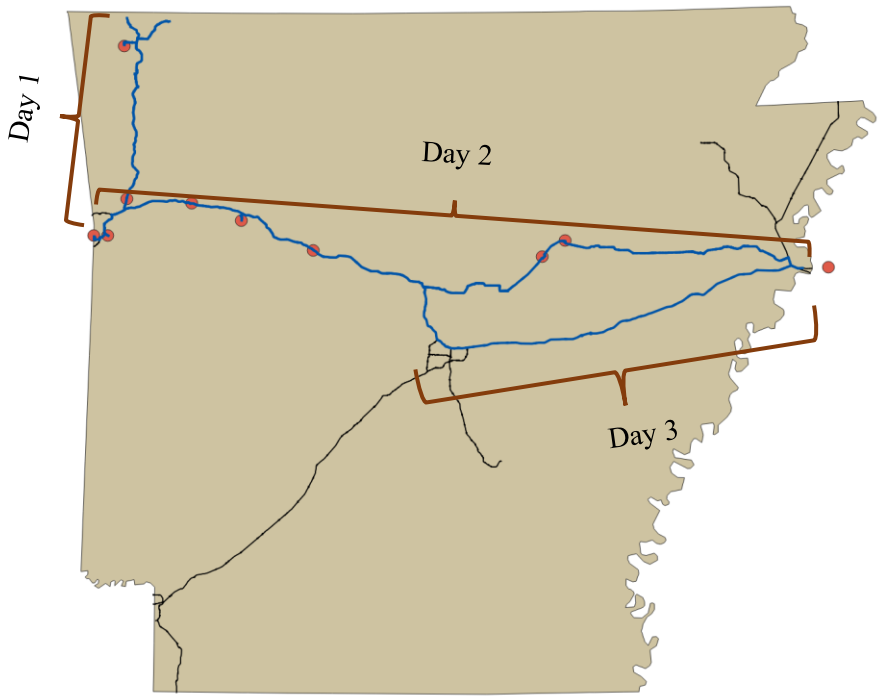


Figure 3.3 Extraction of daily truck movements

Thus, each truck record was represented by an 11-element feature vector based on operational characteristics (Table 3.1). The 11-element features were assumed to distinguish different operational characteristics. For instance, stops of less than 30 minutes duration captured short-breaks (e.g., food break, restroom, and refueling) while stops of 30 minutes to 8 hours

duration captured pickup/delivery stops but not long rest periods (Jing, 2018). Trip length and trip duration were also used to identify the type of truck trips. Trip lengths less than 30 miles and/or trip duration less than 1 hour were assumed to represent short-haul truck movements while trip lengths more than 100 miles and/or trip duration more than 4 hours represented long-haul truck movements.

Table 3.1 Features Defined by Operational Characteristics by Group and Type

Feature Group	Features	Variable Type
Stop Duration	12. Number of stops less than 30 minutes 13. 30 minutes to 8 hours 14. More than 8 hours	Discrete
Trip Length	15. Number of trips less than 30 miles 16. 30 miles to 100 miles 17. More than 100 miles	Discrete
Trip Duration	18. Number of trips less than 1 hour 19. 1 hour to 4 hours 20. More than 4 hours	Discrete
Time of Day (TOD)	21. Proportion of daytime stops (6 AM to 6 PM) to all stops 22. Proportion of nighttime stops (12 AM to 6 AM and 6 PM to 12 AM) to all stops	Continuous

A challenge associated with the business location data is that locations of businesses are reported as the street address location and not the centroid of the building or the truck loading dock. This means that a simple one to one mapping of a truck's stop location to the closest business may not be possible or accurate. For instance, the red dot in Figure 3.4 shows the stop location of a truck within a distribution center (e.g., Walmart Distribution). Calculating the straight-line distance, we found that the distance to the stop was 700 feet from the distribution center and 300 feet from the durable manufacturing store (e.g., Construction of Building Materials). Although the durable manufacturing store was found to be the nearest business location within the ESRI data to that stop, it was not the industry served by that truck since it was visible that the truck was oriented toward the loading area of Walmart (Figure 3.4).

Therefore, instead of assigning a single business/industry to each stop, we developed a probabilistic approach. For each stop made by a truck, an "industry probability matrix" of 31 business categories was estimated based on the presence of each business within a specified spatial buffer around the stop. After manual inspection, 2,000 feet was found to be a suitable buffer distance. Any business establishment found within the buffer distance of a stop was considered equally probable, e.g., if a business establishment of type *b* was found within a buffer distance of 2,000 feet of a stop, we assigned *1* in the probability matrix for business type *b* and otherwise *0*. We did not aggregate probabilities if more than one business establishment of the same type was found in the buffer. For example, if *n* business locations were found within 2,000 feet buffer of a stop where *n-1* were agricultural business and one was a food store, both the agricultural business and the food store would be assigned full probability (i.e., 1). All other business categories would get 0 in the industry probability matrix of that stop.

Using this approach, a probability matrix was estimated for each stop made by a truck. For example, if a truck made three stops, there would be three 31x1 matrices associated with that truck. We assumed that the most frequently visited business type was an indication of the industry served by the truck. For instance, if a truck made three stops, each of which had an mining establishment within its 2,000 ft buffer, then we would assume the truck was serving the mining industry. To identify the likely industry-served by a truck, we combined each 31x1 industry probability matrix for each stop by summing each row, e.g., estimating the total number of stops associated with each industry (Eq. 3.1). We assumed that the industry served by a truck was that with the highest value in the combined matrix.

$$P_i = \sum_{j=1}^{n} [p_{ij}(b_1), p_{ij}(b_2), p_{ij}(b_3), \dots\dots, p_{ij}(b_{31})] \qquad (3.1)$$

Where,

P_i = Total industry probability matrix for truck *i*

p_{ij} = Industry probability matrix for stop *j* of truck *i*

n = Total number of stops for truck *i* in a day

$[b_{1 \ldots} b_{31}]$= 31-NAICS coded business categories (see list in Section 3.4.1)

Business Location	Probability
Livestock Farm	0
Wholesale Trade	1
Restaurant	1
…	…
…	…
…	…
Construction	1
Machinery	0

Figure 3.4 Probability matrix using the proximity analysis

3.4.3 Supervised machine learning for industry classification

Supervised machine-learning is a computer programming method that uses sample data from past experiences to optimize the learning of new experiences in order to satisfy a performance criterion (Alpaydin, 2014). It is often used for classification applications. There are many supervised machine learning tools, including SVMs, neural networks, Bayesian networks, and decision trees (Alpaydin, 2014). Since the "No Free Lunch Theorem" suggests that there is no universally best learning algorithm, the selection of an appropriate model depends on the type of input data and features (Caruana & Niculescu-Mizil, 2006). Breiman (2001) found that

random forest models produce good results in classification with random inputs and random features. Caruana and Niculescu-Mizil (2006) conducted a large-scale empirical comparison between ten supervised learning methods and found that a random forest model was the second-best learning algorithm after calibrated boosted trees. Further, random forest models act to reduce bias and do not overfit to training data (Breiman, 2001). Hence, we selected a Random Forest (RF) model in this work.

RF is an ensemble method that consists of a multitude (or forest) of decision trees from which the final decision (output class) is the average or mode of the classes predicted by the individual trees (Figure 3.5) (Kwok & Carter, 1990).

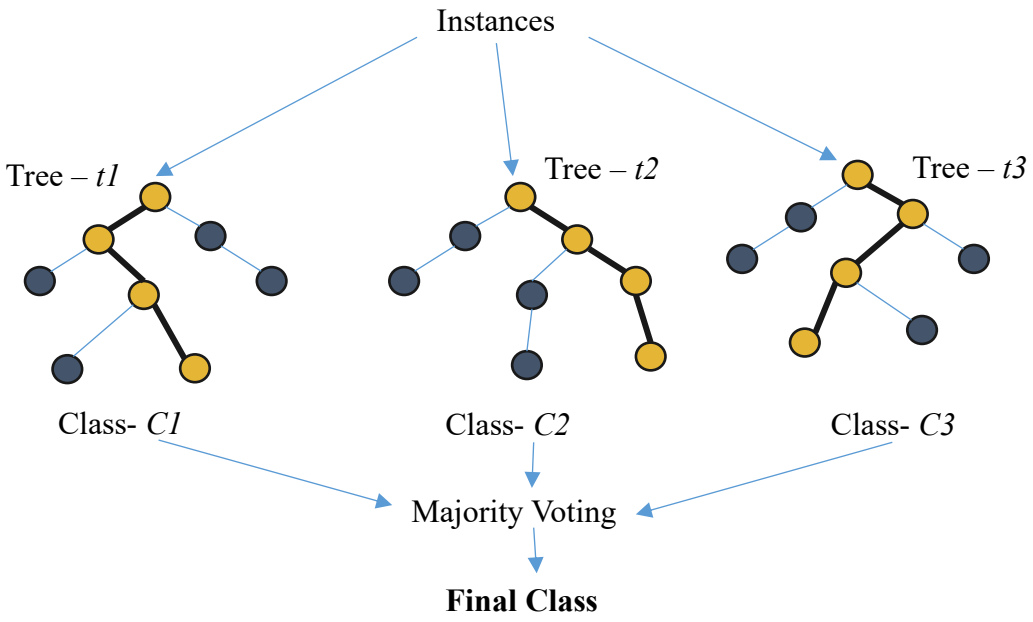


Figure 3.5 A simplified random forest model

Each individual decision tree is a randomized variant of the tree induction algorithm, e.g., each has a randomized "root". Thus, averaging multiple decision trees with different structures consistently produces better results than any of the constituents of the ensemble (Kwok & Carter, 1990). Decision trees are ideal candidates for ensemble methods since they usually have low bias

and high variance, making them highly likely to benefit from the averaging/combining process (Louppe, 2014).

A collection of tree-structured classifiers, $\{h(\boldsymbol{x}, \Theta_k), k = 1,...\}$, makes up a random forest classifier where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class for input $\boldsymbol{x}$ (Breiman, 2001). The formulation for selecting the most popular class for input $\boldsymbol{x}$ is as follows (Biau & Scornet, 2016):

$$\widehat{m}(x) = \sum_{j=1}^{N} \bar{Y}_j \, I\left(x \in A_j\right) \tag{3.2}$$

Where,

$\widehat{m}(x)$ = Majority vote for output class Y for input variable x

A_j = Partition element that contains x

$\bar{Y}_j$ = $n_j^{-1} \sum_{i=1}^{n} \bar{Y}_i \, I\left(x_i \in A_j\right)$, the average of the Y_i's in A_j

n_j = Number of $\{x_i \in A_j\}$

RF implementations mostly differ from each other in the way they introduce random perturbations into the induction procedure (Louppe, 2014). For instance, if our model has two input variables, *X1* = stop duration and *X2* = trip length, a simple classification tree would classify the industry served using these variables (Figure 3.6). Figure 3.6 shows that using stop duration as the root of the tree, if trucks have stop duration < 30 minutes, the model will classify those as *mining* trucks. Next, the tree branches based on trip length, if a truck has stop duration ≥ 30 minutes, the model will check its trip length. If trip length of that truck is < 30 miles, the model will classify that truck as a *mining* truck. The random forest model consists of random variations for the selection of the root note and branches.

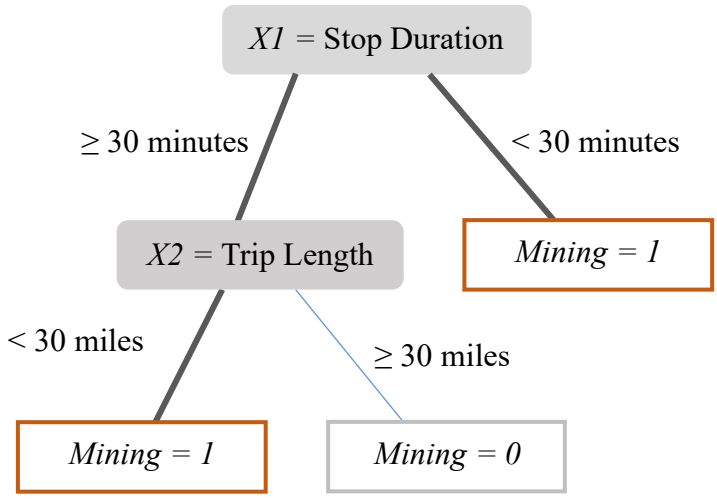


Figure 3.6 A simple classification tree of the model

Supervised machine learning requires labeled training data for model estimation. To generate labeled training data, we compared truck GPS stop locations against aerial imagery of land use and business locations (e.g., Google Satellite images) (Figure 3.7). In contrast to the method proposed in Section 3.4.2 for determining industry served using only business location information (address or lat/long), the satellite imagery allowed us to view building locations, orientations, access roads, loading docks, and other details that provide insight into which business the truck was actually visiting.

Figure 3.7 Stop location of a truck with land use layers and point of interests

Through a sequential inspection, we were able to visually identify the likely industry served by a truck (Figure 3.8). For instance, if a truck had *n* number of stops in a day where each stop was in a gas station, we labeled that truck as an oil and gas truck. In another example, if a truck had *n* number stops in a day where one stop was in a gas station and *n-1* stops were in wholesale distribution centers, that truck was assumed to carry manufacturing goods.

Further, when we found a truck with equal number of stops in multiple unique business locations, we compare the length of time spent at each stop to deduce if the stop was related to the industry served by the truck or used for fuel/rest. For example, if a truck had *n* number of daily stops where *n/2* stops were at distribution centers and *n/2* stops were at gas stations, first, we checked stop duration at each business location. If we found that the truck stopped at distribution centers for t_1 hours and gas stations for t_2 hours where $t_1 > t_2$, we labeled that truck as *manufacturing* truck. Alternatively, if we found that $t_1 = t_2$, we checked the stop time of day at each business location and attempted to label the truck. If we found that it stopped at distribution centers during daytime (6AM-6PM) and at gas stations during nighttime, we assumed that the truck stopped at gas stations for fuel/rest and labeled it as a *manufacturing* truck. Lastly, if the industry served of that truck was still unclear, we checked the trip length and trip duration before making those stops. If the manual inspection found that the trip lengths and durations were longer for distribution centers, we labeled that truck as a *manufacturing* truck. Finally, if we could not identify the industry served of that truck, we labeled that as a "unclassified" truck. Overall, we labeled 2,064 daily truck records according to six distinct industry types (Table 3.2).

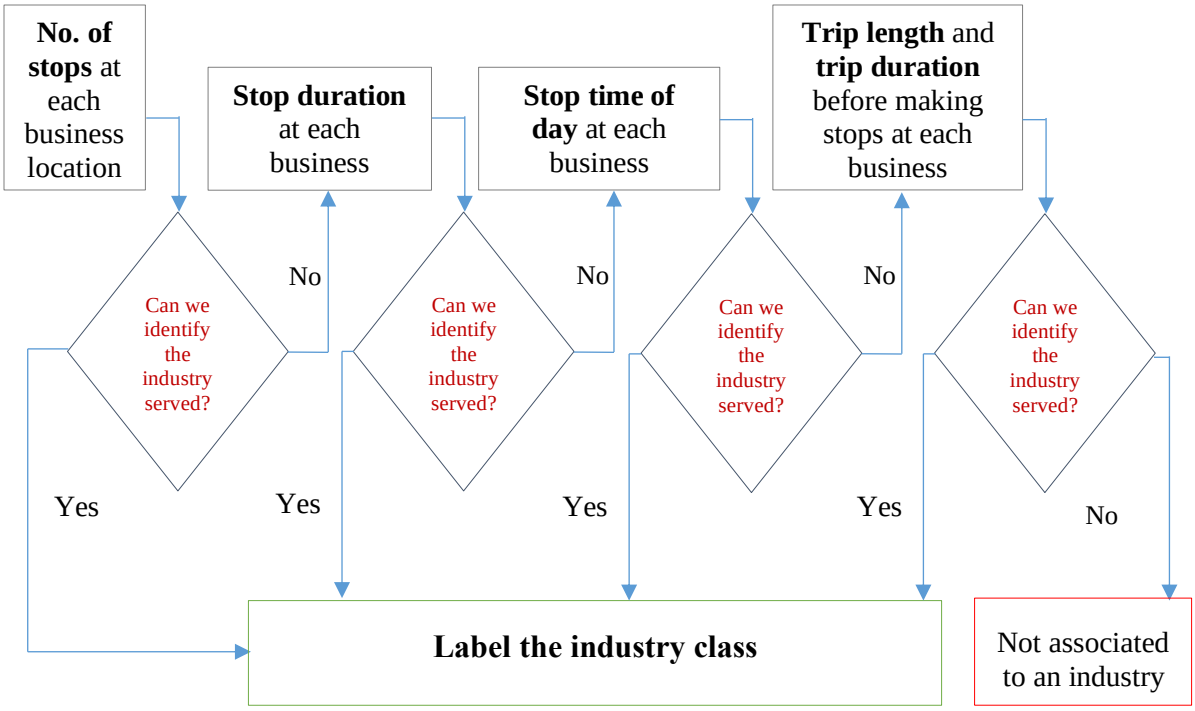


Figure 3.8 Sequential steps to generate labeled training data

Manufactured goods included textiles, food, furniture, plastics, machinery, and equipment) and were labeled as *manufacturing*. Agriculture, forest products, fish, and livestock were included in *farm-products*. *Mining* represented industries related to oil and gas, petroleum, non-metallic minerals, and coal extraction. Chemical industries were grouped into *chemicals* while industries related to clay, concrete, glass, waste, hazardous materials, and small package shipments were grouped into *miscellaneous mixed* class. Lastly, trucks that did not have any industry association within the geographic extent of the study but took fuel break and/or long-rest break were grouped into *pass-through* class.

Table 3.2 Industry Classes Included in the Random Forest Classification Model

Industry Class	Primary Stops' Business Locations
1. Manufacturing	▪ Durable manufacturing ▪ Non-durable manufacturing ▪ Consumer manufacturing ▪ Food manufacturing
2. Farm Products	▪ Agriculture business ▪ Chicken house ▪ Cattle farms ▪ Forests
3. Mining	▪ Gas stations ▪ Gas and oil wells ▪ Gravel field ▪ Mining field
4. Chemicals	▪ Chemical factory and plants ▪ Paint industry ▪ Plastic industry ▪ Rubber industry
5. Miscellaneous Mixed	▪ Shopping malls ▪ Clothing ▪ Accessories
6. Pass-Through	▪ Rest areas ▪ Parking locations ▪ Gas stations ▪ Hotels

3.5 Results

We developed the industry classification model by splitting our input data into a 66/34 training/testing set (Table 3.3). We used the Waikato Environment for Knowledge Analysis (WEKA) framework to develop our model. WEKA uses the "Random Forest" algorithm developed by Leo Breiman and Adele Cutler for inducing a random forest (Kalmegh, 2015). The learning process of the RF classifier followed four steps in our model (Amrehn, Mualla, Angelopoulou, Steidl, & Maier, 2018). First, we drew bootstrap samples $\boldsymbol{B_i}$ for every tree $\boldsymbol{t_i}$ by randomly selecting instances with replacement from $\boldsymbol{X}$ until the sizes of $\boldsymbol{B_i}$ and $\boldsymbol{X}$ were equal. Next, we selected a random subset of features for each $\boldsymbol{B_i}$ and used that as the training of tree $\boldsymbol{t_i}$ in

the forest. Later, we grew unpruned decision trees using bagging mechanism that selected a small subset of features for the split. Finally, a majority vote of the outputs from the individual tree predictions was computed as the final classification result (Table 3.4).

Table 3.3 Distribution of Input Data

Industry Class	Total Instances (% of total sample)	Training Set	Testing Set
Manufacturing	844 (41%)	559	285
Farm Products	501 (24%)	326	175
Mining	626 (30%)	414	212
Chemicals	31 (1.5%)	22	9
Miscellaneous Mixed	35 (1.7%)	26	9
Pass-through	27 (1.3%)	15	12
Total	2,064	1,362	702

To evaluate our classification model, we used performance metrics such as classification accuracy, true positive (TP), true negative (TN), false positive (FP), false negative (FN), and receiver operating characteristic (ROC) area. Classification accuracy (*A*) is the ratio of number of correct predictions to the total number of input samples (Eq. 3.3). Further, when the model correctly predicts the positive class, it is called as true positive. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class while a false negative is an outcome where the model incorrectly predicts the negative class. The false positive rate (*fpr*) and true positive rate (*tpr*) can be calculated using TP, TN, FP, and FN (Eq. 3.4 and 3.5).

$$A = \frac{C}{T} \tag{3.3}$$

$$tpr = \frac{TP}{TP + FN} \tag{3.4}$$

$$fpr = \frac{FP}{FP + TN} \tag{3.5}$$

Where,

A = Classification accuracy

C = Number of correct predictions

T = Total number of predictions made

tpr = True positive rate

TP = True positive

FN = False negative

fpr = False positive rate

FP = False positive

TN = True negative

ROC curves are created by plotting the false positive rate (*fpr*) on the X-axis and the true positive rate (*tpr*) on the Y-axis for each output class of the model (Figure 3.9). The area under the ROC curve can be used to assess the performance of the model. For example, a "steeper" ROC curve indicates low false positive rates and high true positive rates, or higher classification accuracy, for which the ROC area is closer to one (Grzybowski & Younger, 1997). A "shallow" ROC curve indicates an equal number of false and true positive rates, or lower classification accuracy, for which the ROC area is closer to zero.

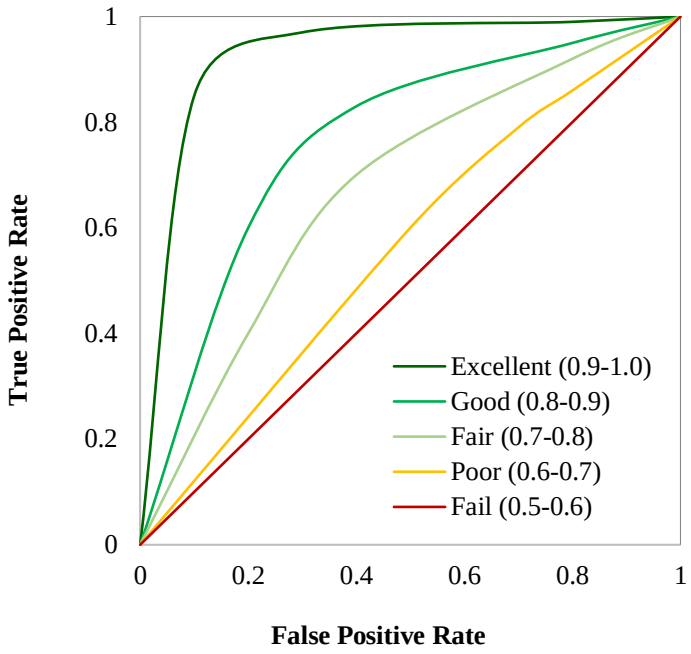


Figure 3.9 Comparison of area under ROC curves

While the training data was used to create the RF classification model, the testing data was used to independently assess model performance. The RF industry classification model predicts six industry classes with an overall Classification accuracy (*A*) of 90% and an overall ROC area of 0.97 (Table 3.4). From the ROC reference curves, we conclude that the overall performance of our model was "excellent" (Figure 3.9). The confusion matrix (Table 3.5) shows the common misclassifications as well as class-specific classification accuracy.

Table 3.4 True Positive and False Positive Rates for Classification Model

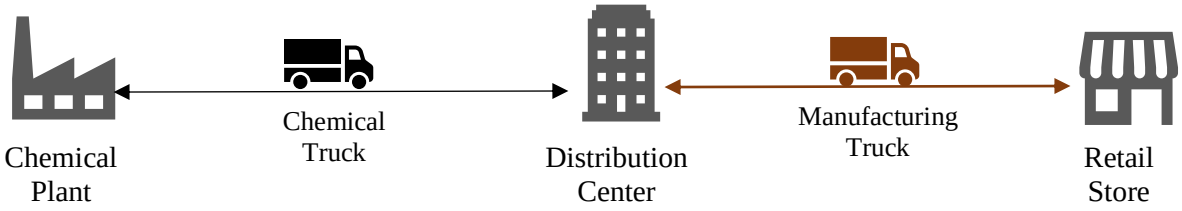
Industry Class	True Positive (TP) Rate	False Positive (FP) Rate	ROC Area
Manufacturing	0.96	0.13	0.97
Farm Products	0.89	0.01	0.99
Mining	0.86	0.03	0.98
Chemicals	0.67	0.00	0.93
Miscellaneous Mixed	0.67	0.00	0.95
Pass-Through	0.50	0.00	0.87
Weighted Average	**0.90**	**0.06**	**0.97**

Table 3.5 Confusion Matrix of the Classification Model

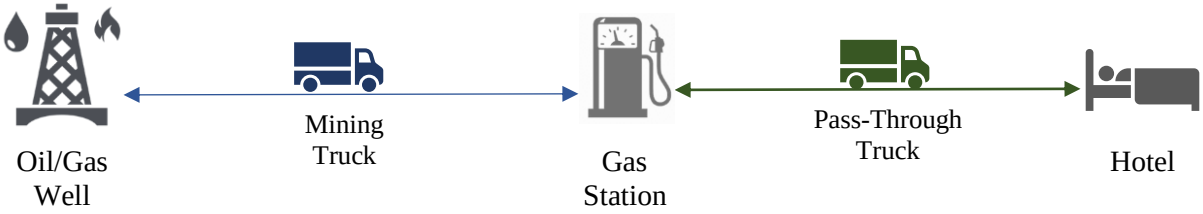
Actual Instances	Classified As						Classification Accuracy
	Manufact.	Farm Products	Mining	Chemicals	Misc. Mixed	Pass-through	
Manufacturing	**274**	1	10	0	0	0	**96%**
Farm Products	18	**155**	2	0	0	0	**89%**
Mining	28	2	**182**	0	0	0	**86%**
Chemicals	3	0	0	**6**	0	0	**67%**
Misc. Mixed	3	0	0	0	**6**	0	**67%**
Pass-through	3	0	3	0	0	**6**	**50%**

Manufacturing had the highest classification accuracy but also the highest false positive rate (0.13). This indicates that model incorrectly classified other industry classes as manufacturing. As a result, *chemicals* and *miscellaneous mixed* industry classes had lower classification accuracy rates (around 67%) since the model incorrectly classified these classes as *manufacturing*. This is likely the result of similarities in operating characteristics among these industry classes. For instance, trucks serving chemical industries such as plastic industries commonly made stops for delivery of plastic bags and packaging materials at manufacturing distribution centers while manufacturing trucks made stops there for pickup consumer packaged goods (Figure 3.10a). Since these trucks shared one common business location, there was a possibility of misclassification when their other business locations also coincided. In this example, our model would misclassify those trucks if the chemical plant and retail store both were within 2,000 feet buffer distance of stops (Figure 3.10a). Likewise, miscellaneous mixed trucks were misclassified as manufacturing trucks. Another common misclassification was that of pass-through trucks as mining. Similar to the example of *chemical* and *manufacturing* trucks (Figure 3.10a), Figure 3.10b shows that *mining* and *pass-through* trucks share a common business location, gas stations. Besides stops at gas stations, *mining* trucks made stop at oil/gas wells while *pass-through* trucks, occasionally, made another stop at hotels or rest areas for long-

rest (Figure 3.10b). If oil/gas well and hotel both were within 2,000 feet buffer distance of stops, our model would misclassify *pass-through* trucks as *mining* trucks.



(a) Example of *chemical* and *manufacturing* trucks



(b) Example of *mining* and *pass-through* trucks

Figure 3.10 Industry-specific truck stops at different business locations

Imbalance data sets can degrade the performance of machine learning models since decisions may be biased toward the majority classes (Elrahman & Abraham, 2013). This leads to the common misclassification in the minority classes. This challenge is known as "rare event detection" or the "class imbalance problem". In our model, we have three minority classes with low number of training data samples including 22 trucks for *chemicals*, 26 trucks for *miscellaneous mixed*, and 15 trucks for *pass-through* industries (Table 3.3). These three minority classes were also produced low accuracy rates compared to other classes (Table 3.5). As a solution to this problem, we propose *over sampling methods* that suggest increasing the *groundtruth* data for minority classes to improve the accuracy rate for these classes (Elrahman & Abraham, 2013). Although *under sampling methods* can also handle "class imbalance problem", it may cause loss of useful information by removing significant patterns (Elrahman & Abraham,

2013). Hence, we were inclined to use *over sampling methods* to improve the performance of the model. Further, varying the size of the buffer based on the density of business in an area can be another solution. Instead of using a fixed probability of 1 for any business in the buffer, we propose changing the probability relative to the distance in future work.

3.6 Discussion

The determination of ideal training/testing split is a challenge in developing a classification model as the small ratio of training data may cause loss of useful information and the large ratio may cause overfitting (Elrahman & Abraham, 2013). To demonstrate the sensitivity of our classification model, we examined two split ratios of training/testing data sets, i.e., 55/45 and 85/15. In this section, we present 55/45 ratio as "under sample", 85/15 ratio as "over sample", and 66/34 ratio as "base sample". We compared the results of the new two models with our classification model (Table 3.6).

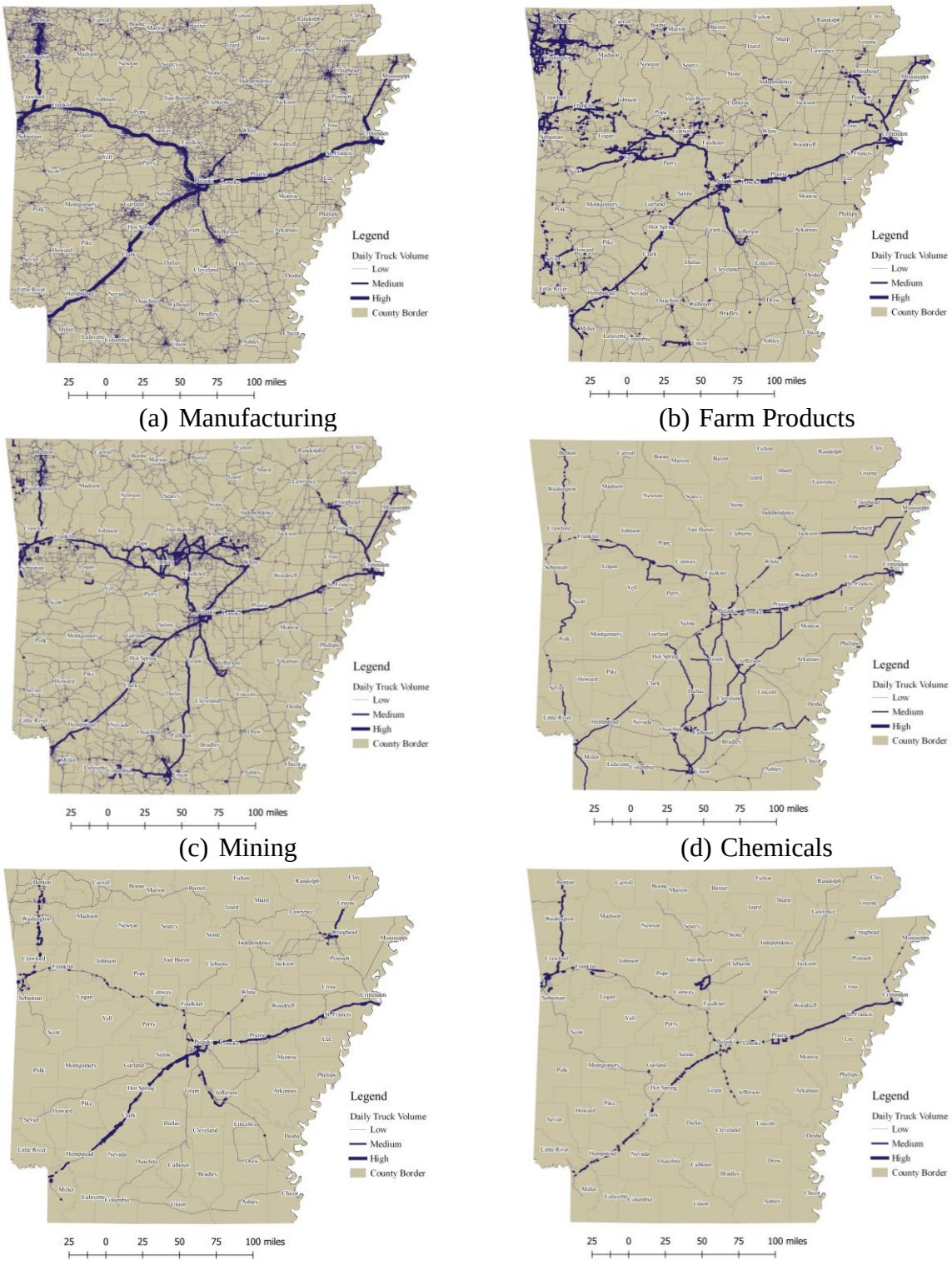
Table 3.6 A Comparison of Classification Accuracy for Different Training/Testing Ratio

Industry Class	**Classification Accuracy**		
	Base Sample (66/34 Ratio)	**Under Sample** (55/45 Ratio)	**Over Sample** (85/15 Ratio)
Manufacturing	96%	96%	99%
Farm Products	89%	89%	91%
Mining	86%	84%	90%
Chemicals	67%	67%	50%
Misc. Mixed	67%	46%	33%
Pass-through	50%	41%	25%

We found from the comparison table that the model developed with "under sample" could not improve the accuracy rates for any industry classes but degraded for *mining, miscellaneous mixed,* and *pass-through* industries compared to the "base sample" model (Table 3.6). Unlike "under sample" model, "over sample" model could improve the accuracy rates for

the majority classes (e.g., *manufacturing*, *farm products*, and *mining*) (Table 3.6). However, the "over sample" model degraded the accuracy rates for the minority classes (e.g., *chemicals*, *miscellaneous mixed*, and *pass-through*) (Table 3.6). Since over fitting of data may cause the improvement for the majority classes but drop for the minority classes (Elrahman & Abraham, 2013), we suggested not to develop the model with "over sample" data. Similarly, we assumed that the "base sample" model was the better model for industry classification (Table 3.6).

Unlike survey data commonly used to understand commodity flows across a state region, GPS data allows us to see complete paths between Origins and Destinations (ODs). By applying the RF method to predict industry from GPS data, we gain valuable insight in the OD flows by link and by industry. Since origin-destination (OD) flows differ by industry and commodity, we observed distinct truck paths for each industry class (Figure 3.11). For instance, we can see that trucks serving the manufacturing industry rely heavily on the interstate system but are distributed across the entire state (Figure 3.11a). Trucks carrying *farm products*, on the other hand, were highly concentrated near crop fields, chicken houses, cattle farms, and forests (Figure 3.11b). *Mining* trucks showed a high concentration where oil/gas wells were located, e.g., Conway county. Since those trucks made frequent stops at gas stations, they were found all over the state (Figure 3.11c). Truck carrying *chemicals* were concentrated in southern part of the state where several chemical plants were located (Figure 3.11d). The model also predicted a small number of trucks as *miscellaneous mixed* (Figure 3.11e) and *pass-through* movements (Figure 3.11f). As per our definition, *pass-through* trucks crossed over the state were not associated to any industry. Typically, those trucks made stops at gas stations and/or rest areas. Therefore, most of those trucks were seen on the interstates and highways (Figure 3.11f).



(a) Manufacturing (b) Farm Products

(c) Mining (d) Chemicals

(e) Miscellaneous Mixed (f) Pass-Through

Figure 3.11 Truck volumes on roads for different industry class

The RF classification model can predict the industry class for a given truck based on that truck's daily activity pattern represented by its number of daily stops, stop location, stop duration, trip length, trip duration, and an estimated probability of the businesses visited at each stop. To train and test the model, we applied it to daily activity patterns from a statewide sample of trucks which had been manually labeled according to their industry served. At present, the manually labeled data is the only source of data linking daily operational characteristics to industry or commodity. However, we recognize that it is endogenous to our model development.

As a means to validate the model using data independent from the GPS samples, we compared the volumes of trucks by industry estimated from our model to the Arkansas Statewide Travel Demand Model (ARSTDM) (ARDOT, 2012). The purpose of this comparison is not for direct validation of our model, since our model and the ARSTDM differ in trip definitions, data sources, time periods, etc., but rather to provide context for our model's contributions.

For this comparison, the trained RF classification model was applied to 278,990 daily truck movement records. Since truck GPS data is a sample of the total truck population, we expanded the sample to represent the entirety of the truck population. Expansion factors were derived by comparing the GPS volumes to tuck traffic volumes measured by Weigh-in-Motion (WIM) sensors. On average, the statewide sample of GPS data in Arkansas represented 10-15% of the total truck traffic (Akter, Hernandez, Diaz, & Ngo, 2018; Corro, Akter, & Hernandez, 2019).

Input data for the ARSTDM was collected from TRANSEARCH, a proprietary commodity flow database. TRANSEARCH amalgamates a variety of survey datasets including the national CFS but the procedure to combine multiple datasets and the datasets themselves are not disclosed. The ability to replace or supplement TRANSEARCH data with observed GPS

data has the potential to improve ARSTDM accuracy. The ARSTDM contains predictions for 15 commodity groups. Thus, it was necessary to link our five industry groups to the 15 commodity groups (Table 3.7). The base year of the ARSTDM was 2010 while our industry predictions were derived from truck GPS data from 2016. No attempt was made to bring the datasets into the same time period.

Table 3.7 Linking Industry Class to ARSTDM Commodity Groups

Industry Class used in RF Model	ARSTDM			Total Tonnage by Industry Class (% of total)
	Commodity Groups	Tons by Trucks	Average Payload factor (Tons/Truck)	
Manufacturing	Food	34,553,853	23.0	120,505,775 (39%)
	Consumer Manufacturing	2,758,042	18.64	
	Non-Durable Manufacturing	13,199,197	12.7	
	Lumber	42,858,920	25.3	
	Durable Manufacturing	27,135,763	15.78	
Farm Products	Farm Products	19,416,929	16.26	19,416,929 (6%)
Mining	Mining/ Metallic ores	1,662,389	22.64	69,866,693 (23%)
	Coal	109,006	24.81	
	Nonmetallic Minerals	52,599,184	24.31	
	Petroleum	15,496,114	24.07	
Chemicals	Chemicals	14,019,807	20.67	14,019,807 (5%)
Miscellaneous Mixed	Paper	5,176,079	24.04	86,129,669 (28%)
	Clay, Concrete, Glass	28,983,834	17.17	
	Primary Metal	14,549,205	24.88	
	Secondary & Misc. Mixed	37,420,551	20.56	
Pass-through	N/A	N/A	N/A	

We calculated truck volume for each industry class by multiplying payload factors by their respective tonnages and then summing the volumes of each commodity group (Table 3.7). We compared the percent of total truck volume between the ARSTDM and our RF classification method instead of total truck volume (Figure 3.12). Overall, the magnitude of truck volumes by industry estimated by ARSTDM and our RF classification model are in general agreement. The percentage estimated from the ARSTDM and our RF model of trucks carrying *farm products* and *mining* were similar. It is expected as our model can predict *farm products* and *mining* with high accuracy and precision, e.g., ROC areas of 0.99 and 0.98, respectively (Table 3.4). In reference to model performance, *manufacturing* has the highest false positive rate (0.13) which results in an overestimation of truck volume for that class (Figure 3.12). Similarly, *chemicals* and *miscellaneous mixed* have lower accuracy rate (67%) in our model and hence, GPS truck volume for these two industries show larger gap with ARSTDM truck volume.

Other discrepancies may be caused by inaccurate conversions and/or different definitions of trips within the two datasets. Since ARSTDM used TRANSEARCH commodity flow data, it was necessary to use payload factors to convert tonnage flow to truck volumes. It is possible that truck volumes could be underestimated, possibly, for industries that have a higher prevalence of empty haul and less than truck load (LTL) movements. For example, if we assume that the number of empty haul/ LTL truck movements is higher for manufacturing (and thus the payload factor is actually lower that shown in Table 3.7) compared to farm products and mining industry, then the discrepancy between the ARSTDM and RF Classification results can be accounted for.

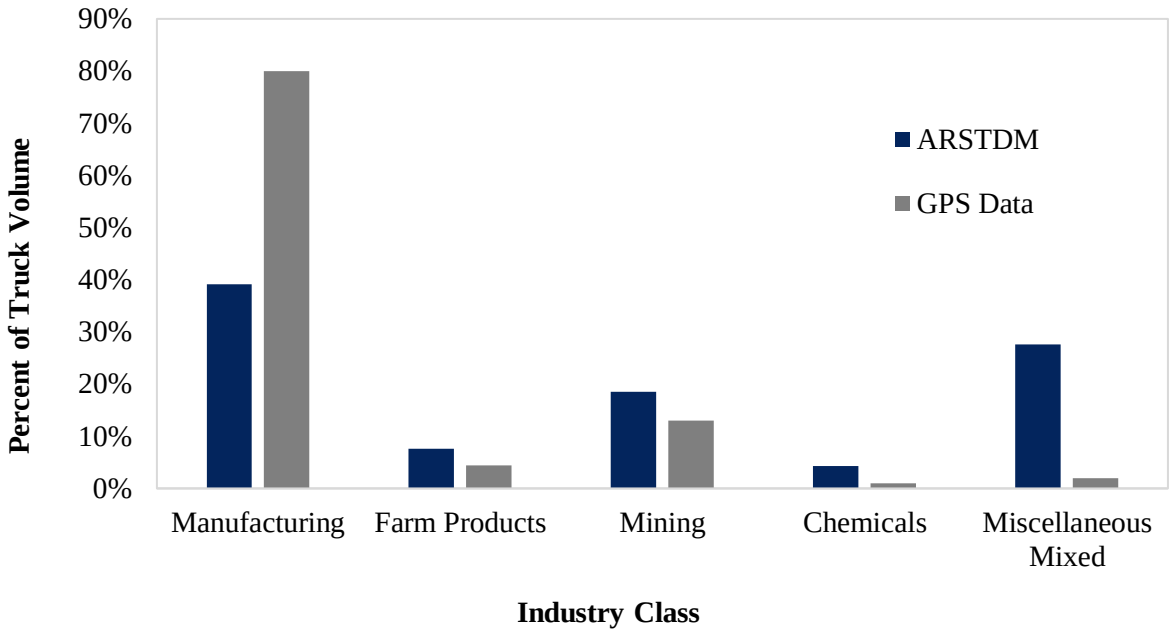


Figure 3.12 Percentage of truck volume by industry class in ARSTDM and GPS data

3.7 Conclusions

Freight travel demand forecasting models estimate future road usage patterns by first predicting economic growth by industry sector. These models require data relating truck operational characteristics to industry served. Knowledge of the connection between a truck's operational characteristics and the commodity it carries or the industry it serves can provide insight into road usage patterns between origins and destinations (Beagan, Tempesta, & Proussaloglou, 2019). However, this data is not available from state-of-the-practice data sources like driver, shipper, and carrier surveys. To address this data need, we developed a method to predict industry served from mobile sensor data, specifically GPS data. GPS data represents an increasingly available source of big data for freight that reveals the position of trucks over time and space and has almost ubiquitous network coverage. But due to data sharing restrictions, this private sector data source is shared with the public sector only after sterilizing identifiable

information like industry served. Therefore, our approach is necessary to re-identify industry served while maintaining privacy standards.

The premise of our model is that advanced pattern recognition tools, i.e., supervised machine learning techniques, can predict industry served by a truck based on inputs related to the truck's daily activity pattern. These activity patterns depict trip chains detailing stop and journey sequences. Using a probabilistic method to assign a likely industry type to each stop in a trip chain, we are able to predict among six industry classes with 90% accuracy using a Random Forest (RF) supervised learning model. Over 2,064 daily truck records were used for model training and testing. Further, as a means of validation the RF model was applied to over 270,000 daily truck records and truck volumes by industry class estimated by the RF model were compared a statewide commodity inventory, e.g. the input commodity tonnages from a statewide freight demand forecasting model. Despite differences in data collection methods, time periods, and trip definitions, the magnitude of truck volumes by industry estimated by the statewide model and our RF classification model are in general agreement.

The RF classification model can predict six distinct industry classes that represent 15 aggregated commodity groups. Although commodities were aggregated by industry sector, aggregation may be responsible for lower classification accuracy. For instance, *manufacturing* included four industries that produce varied commodities such as furniture, electrical equipment, machinery, food products, etc. It is possible that each of these four industries differs in its operational characteristics like stops per day. Hence, to improve the model we will disaggregate industry classes.

Disaggregation may, however, be limited by our ability to generate sizeable labeled samples for model training and testing. The training data used to develop this model was

imbalanced as *chemicals, miscellaneous mixed,* and *pass-through* industry classes had less than 30 samples, e.g., minority classes. Alternatively, the *manufacturing*, *mining,* and *farm products* industries were considered as majority classes with more than 300 samples. To solve this class imbalance problem, we suggest increasing the training data for the minority classes, or as previously mentioned oversampling the minority classes during training.

Additionally, we observed that the buffer distance used to create our industry probability matrix, which represented the probability of each of 31 business types within a 2,000ft buffer of the truck's stop, may contribute to misclassifications. The 2,000 ft buffer was selected by trial and error by comparing model accuracy with changes to the buffer size. To improve this method, we could vary the size of the buffer based on the density of business in an area. We also propose changing the probability relative to the distance from the truck's stop to improve the performance of the model.

Ultimately, our developed model demonstrated that operational characteristics of trucks, i.e, the number of stops, stop location, stop duration, stop time of day, trip length, and trip duration have distinct patterns based on commodity carried and industry served.

3.8 Acknowledgement

The authors thank the Arkansas Department of Transportation (ARDOT) for sponsoring the project that led to this paper.

3.9 Authors Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: T. Akter and S. Hernandez; data gathering and processing: T. Akter; analysis and interpretation of results: T. Akter and S. Hernandez; draft manuscript preparation: T. Akter. All authors reviewed the results and approved the final version of the manuscript.

3.10 References

Adler, T., & Ben-Akiva, M. (1979). A Theoretical and Empirical Model of Trip Chaining Behavior. *Transportation Research Part B, 13*(3), 243-257. doi:10.1016/0191-2615(79)90016-X.

Akter, T., Hernandez, S., Diaz, K. C., & Ngo, C. (2018). Leveraging Open-Source GIS Tools to Determine Freight Activity Patterns from Anonymous GPS Data. Paper presented at the *AASHTO GIS for Transportation Symposium.*

Allahviranloo, M., & Recker, W. (2013). *Daily Activity Pattern Recognition by Using Support Vector Machines with Multiple Classes* doi:https://doi.org/10.1016/j.trb.2013.09.008.

Allahviranloo, M., Regue, R., & Recker, W. (2017). Modeling The Activity Profiles of a Population. *Transportmetrica B: Transport Dynamics, 5*(4), 426-449. doi:10.1080/21680566.2016.1241960.

Alpaydın, E. (2014). *Introduction to Machine Learning* (3. ed. ed.). Cambridge, Mass. [u.a.]: MIT Press. Retrieved from http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=027423356&sequence=000005&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Amrehn, M., Mualla, F., Angelopoulou, E., Steidl, S., & Maier, A. (2018). *The Random Forest Classifier in WEKA: Discussion and New Developments for Imbalanced Data.*

ARDOT. (2012). *Arkansas Statewide Travel Demand Model*.

ATRI. (2019). Freight Performance Measures. Retrieved from https://truckingresearch.org/tag/freight-performance-measures/.

Bassok, A., McCormack, E. D., Outwater, M. L., & Ta, C. (2011). Use of Truck GPS Data for Freight Forecasting. Paper presented at the *Transportation Research Board 90th Annual Meeting.*

Beagan, D., Tempesta, D., & Proussaloglou, K. (2019). *Quick Response Freight Methods.* (). Retrieved from https://ops.fhwa.dot.gov/publications/fhwahop19057/fhwahop19057.pdf.

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *Test, 25*(2), 197-227.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324.

Camargo, P., Hong, S., & Livshits, V. (2017). Expanding The Uses of Truck GPS Data In Freight Modeling And Planning Activities. *Transportation Research Record, 2646*(1), 68-76. doi:10.3141/2646-08.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. Paper presented at the 161-168. doi:10.1145/1143844.1143865 Retrieved from http://dl.acm.org/citation.cfm?id=1143865.

Chow, J., Yang, C., & Regan, A. (2010). State-of-the Art of Freight Forecast Modeling: Lessons Learned and the Road Ahead. *Transportation, 37*(6), 1011-1030. doi:10.1007/s11116-010-9281-1.

Corro, K. D., Akter, T., & Hernandez, S. (2019). Comparison of Overnight Truck Parking Counts with GPS-Derived Counts for Truck Parking Facility Utilization Analysis. *Transportation Research Record, 2673*(8), 377-387. doi:10.1177/0361198119843851.

Elrahman, S. M. A., & Abraham, A. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing, 1*(2013), 332-340.

FHWA. (2014). *All Road Network of Linear Referenced Data (ARNOLD) Reference Manual*. Retrieved from https://www.fhwa.dot.gov/policyinformation/hpms/documents/arnold_reference_manual_2014.pdf.

FHWA. (2018). Status of the Nation's Highways, Bridges, and Transit Conditions and Performance: 23rd Edition: Part III: Highway Freight Transportation - report to congress. Retrieved from https://ops.fhwa.dot.gov/freight/infrastructure/nfn/rptc/cp23hwyfreight/iii_ch11.htm.

FHWA. (2019a). Freight Analysis Framework Version 4. Retrieved from http://faf.ornl.gov/fafweb/.

FHWA. (2019b). Freight Management and Operations - Freight Analysis Framework 3 User Guide. Retrieved from https://ops.fhwa.dot.gov/freight/freight_analysis/faf/faf3/userguide/.

Giovannini, L. (2011). *A Novel Map-Matching Procedure for Low-Sampling GPS Data with Applications to Traffic Flow Analysis* doi:10.6092/unibo/amsdottorato/3898. Retrieved from https://www.openaire.eu/search/publication?articleId=od_____1754::2e76bee797112fda11280f4851def321.

Grzybowski, M., & Younger, J. G. (1997). Statistical Methodology: III. Receiver Operating Characteristic (ROC) Curves. *Academic Emergency Medicine, 4*(8), 818-826. doi:10.1111/j.1553-2712.1997.tb03793.x.

Jiang, S., Ferreira, J., & González, M. (2012). Clustering Daily Patterns of Human Activities in the City. *Data Mining and Knowledge Discovery, 25*(3), 478-510. doi:10.1007/s10618-012-0264-z.

Jing, P. (2018). *Identifying and Modeling Urban Truck Daily Tour-Chaining Patterns.*

Kalmegh, S. R. (2015). Comparative Analysis of Weka Data Mining Algorithm Randomforest, Randomtree and Ladtree for Classification of Indigenous News Data. *International Journal of Emerging Technology and Advanced Engineering, 5*(1), 507-517.

Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Vallabhaneni, L., & Nippani, S. (2014). Development of a Tour-Based Truck Travel Demand Model Using Truck GPS Data. Paper presented at the *93rd Annual Meeting of the Transportation Research Board;*

Kwok, S. W., & Carter, C. (1990). Multiple Decision Trees. *Machine Intelligence and Pattern Recognition, 9,* 327-335.

Li, S., & Lee, D. (2017). Learning Daily Activity Patterns with Probabilistic grammars. *Transportation, 44*(1), 49-68. doi:10.1007/s11116-015-9622-1.

Liao, C. (2009). *Using Archived Truck GPS Data for Freight Performance Analysis on I-94/I-90 from the Twin Cities to Chicago.*University of Minnesota Center for Transportation Studies. Retrieved from https://conservancy.umn.edu/handle/11299/97668.

Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., & Cools, M. (2014). Building a Validation Measure for Activity-Based Transportation Models Based on Mobile Phone Data. *Expert Systems with Applications, 41*(14), 6174-6189. doi:https://doi.org/10.1016/j.eswa.2014.03.054.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice* Retrieved from http://orbi.ulg.ac.be/handle/2268/170309.

Ma, X., McCormack, E. D., & Wang, Y. (2011). Processing Commercial Global Positioning System Data to Develop a Web-Based Truck Performance Measures Program. *Transportation Research Record: Journal of the Transportation Research Board, 2246*(1), 92-100. doi:10.3141/2246-12.

Mortazavi, B. J., Downing, N. S., Bucholz, E. M., Dharmarajan, K., Manhapra, A., Li, S. X., ... & Krumholz, H. M. (2016). Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circulation: Cardiovascular Quality and Outcomes*, *9*(6), 629-640. doi:10.1161/CIRCOUTCOMES.116.003039.

Quddus, M., & Washington, S. (2015). Shortest Path and Vehicle Trajectory Aided Map-Matching for Low Frequency GPS Data. *Transportation Research Part C: Emerging Technologies, 55,* 328-339. doi:https://doi.org/10.1016/j.trc.2015.02.017.

Sarti, L., Bravi, L., Sambo, F., Taccari, L., Simoncini, M., Salti, S., & Lori, A. (2017). Stop Purpose Classification from GPS Data of Commercial Vehicle Fleets. Paper presented at the *2017 IEEE International Conference on Data Mining Workshops (ICDMW),* 280-287.

Sharman, B. W., & Roorda, M. J. (2011). Analysis of Freight Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board, 2246*(1), 83-91. doi:10.3141/2246-11.

Shoval, N., & Isaacson, M. (2007). Tracking Tourists in the Digital Age. *Annals of Tourism Research, 34*(1), 141-159. doi:10.1016/j.annals.2006.07.007.

Sun, Z., & Ban, X. (. (2013). Vehicle Classification Using GPS Data. *Transportation Research Part C: Emerging Technologies, 37*, 102-117. doi:https://doi.org/10.1016/j.trc.2013.09.015.

Thakur, A., Pinjari, A. R., Zanjani, A. B., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board, 2529*(1), 66-73. doi:10.3141/2529-07.

Theja, P V V K, & Vanajakshi, L. (Nov 2010). Short Term Prediction of Traffic Parameters Using Support Vector Machines Technique. Paper presented at the 70-75. doi:10.1109/ICETET.2010.37 Retrieved from https://ieeexplore.ieee.org/document/5698294.

Xie, Y., & Huynh, N. (2010). Kernel-Based Machine Learning Models for Predicting Daily Truck Volume at Seaport Terminals. *Journal of Transportation Engineering, 136*(12), 1145-1152. doi:10.1061/(ASCE)TE.1943-5436.0000186.

Yang, X., Sun, Z., Ban, X. J., & Holguín-Veras, J. (2014). Urban Freight Delivery Stop Identification with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2411*(1), 55-61. doi:10.3141/2411-07.

YANG, Y., YAO, E., YUE, H., & LIU, Y. (2010). Trip Chain's Activity Type Recognition Based On Support Vector Machine. *Journal of Transportation Systems Engineering and Information Technology, 10*(6), 70-75. doi:10.1016/S1570-6672(09)60073-8.

Zanjani, A. B., Pinjari, A. R., Kamali, M., Thakur, A., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2494*(1), 87-96. doi:10.3141/2494-10.

Chapter 4

4 A Spatial Panel Regression Model to Measure the Effect of Weather Events on Freight Truck Traffic

Taslima Akter*, Suman Kumar Mitra, Sarah Hernandez, and Karla Corro-Diaz

*Corresponding author. Email: takter@uark.edu

4.1 Abstract

Truck drivers adhere to delivery schedules making them more likely to reroute rather than cancel a trip when faced with inclement weather. While previous studies modeled the direct effects of adverse weather on total traffic volumes, none considered the particular implications for trucks. The ability to predict spatial and temporal shifts in truck traffic resulting from adverse weather is novel and useful for decision makers tasked with long-range freight planning and for the trucking industry. With deeper insights into rerouting around adverse weather, the trucking industry will be able to more efficiently plan and accurately estimate billable miles. Thus, this study applied dynamic spatial panel regression that captures rerouting behavior of trucks due to adverse weather conditions. Results showed that changes in truck traffic volume due to adverse weather conditions, e.g., surface runoff, snow mass, and humidity, exhibited spatial (direct and indirect) and temporal shifts (short and long term effects).

4.2 Introduction

Trucking is a critical component of the freight transportation system. Although freight shipments traverse a multimodal system comprised of air, rail, pipeline, and truck modes, trucking is and is forecast to be the dominant mode for freight. In 2015, trucks account for 64% and 69% of the market by both weight and value, respectively (FHWA, 2018a). Further, the

Freight Analysis Framework, the FHWA's nationwide freight forecasting model estimates that the weight of freight shipments moved by truck will grow 45% between 2012 and 2045 (FHWA, 2018a). A reliable estimation of truck-based freight travel demand and behavior is necessary for effective planning, design, and management of freight transportation system infrastructure and operations (FHWA, 2019).

Severe weather conditions such as extreme hot and cold temperatures, high wind speeds, icy conditions, and snowfall and snow accumulation, affect traffic volumes along the highway network (Melillo, 2014). Weather events such as tornadoes and flooding can cause significant disruptions to the freight transportation network resulting in economic impacts to the trucking industry and, consequently, industries served by the trucking industry. Such impacts include displaced congestion effects as well as shipment delays, depreciation of goods, and inventory holding costs (Winston & Shirley, 2004). Impacts to or in the vicinity of Primary Freight Network (PFN) segments, in particular, will have far reaching effects on freight movements across the nation. Winston and Shirley (2004) estimated that the annual cost of congestion for a state was around $7 billion. Ivanov et al. (2008) estimated that the total loss from freight delay, due to the storm-related two corridor closures, was almost $75 million. Understanding the impacts of weather events on freight movements can help state agencies better predict the impacts of such events for operational purposes (e.g., detours, traveler information signs, etc.) and as a means to provide more accurate monetized cost/benefit estimates for highway infrastructure maintenance or improvement projects. Moreover, understanding rerouting and delay caused by adverse weather conditions can help identify critical links and improve resiliency measurement and planning. Beyond public sector planning, recognizing and being able to model the effects of adverse weather conditions allows the trucking industry to better plan

routes, estimate time of arrival, and accurately calculate billable and revenue miles. To assess impacts such as route changes and time delays, better models are needed to predict the number of affected vehicles and geographic extent of the impacts.

The impact of adverse weather conditions, such as those resulting from winter storms including snowfall or snow mass (i.e. snow accumulation), can be measured in part by differences in traffic volumes along the transportation network. Faced with adverse weather, drivers may postpone a trip (i.e. temporal shift), change routes (i.e. spatial shift), or cancel a trip all together (i.e. volume reduction) (Datla, Sahu, Roh, & Sharma, 2013). In the presence of adverse weather such as snowstorms, total traffic volumes can reduce as much as 56% (Hanbali & Kuemmel, 1993) since many travelers choose to cancel their trips. However, trucks are less likely to cancel trips due to adverse weather conditions compared to passenger vehicles (Maze et al., 2005). Since freight is subject to rigid pickup/delivery schedules, freight truck drivers have less flexibility in the decision to travel, instead choosing to reroute and/or delay their trip (Winston & Shirley, 2004). Consequently, while reductions in total traffic volumes may occur due to adverse weather conditions, freight truck traffic may actually increase along certain routes (e.g., official or unofficial detours) (Datla, Sahu, Roh, & Sharma, 2013). However, previous studies could not capture the rerouting behavior of trucks using simple linear regression models.

The goal of this study was to develop a predictive model that captures the spatial and temporal rerouting behavior of freight trucks due to adverse weather conditions (e.g., snowfall, rainfall, etc.). The study employed a dynamic spatial panel regression model to predict the percentage change in daily freight truck volume due to adverse weather conditions. The dynamic spatial panel regression model incorporated (i) temporal data including historical truck volume

trends, seasonality, and daily variations in traffic volumes and (ii) variables that capture adverse weather conditions, i.e. average humidity, surface runoff, and snow mass.

Beyond developing a model specifically for truck traffic volume prediction, three novel expansions of the studies described above are presented in this paper: (i) improvements in the spatial and temporal scope and resolution of the traffic data, (ii) expansion of existing modeling techniques to include dynamic spatial panel regression techniques, and (iii) consequent on (i) and (ii), the ability to demonstrate and measure rerouting behavior of freight trucks due to weather conditions. As it relates to (i), six years (2011-2016) of daily truck volume data from 18 Weigh-in-Motion (WIM) stations in Arkansas and corresponding weather data from the Modern-Era Retrospective analysis for Research and Applications (MERRA), a weather dataset provided by the Long Term Pavement Performance (LTPP) InfoPave Climate Tool, were used to develop the model. In this study, daily truck volume across 114 days from each WIM station was used to expand the temporal scope. As it relates to (ii), existing studies fail to capture the spatial and temporal autocorrelation of weather and truck traffic volume within their modeling specifications. This study employed dynamic spatial modeling to accurately model and capture such effects. Spatial diagnosis was performed by first estimating an OLS model to select the appropriate spatial model, e.g., a Spatial Error model or a Spatial Autoregressive model (SAR). Ultimately, a dynamic SAR model with spatial fixed effects was developed to predict the percentage change in truck volume due to weather related variables, day of the week, season, and historical trends in daily truck volumes. As it relates to (iii), the chosen model specification interprets the dependent variable as time-lagged and space-time-lagged. Thus, at any location, the estimated model can predict the percentage change in truck volume resulting from adverse

weather conditions at that location (direct effects), at neighboring locations (indirect effects), at the immediate time periods (short term), and at delayed time periods (long term).

A better understanding of the effect of weather conditions on truck traffic can help state and regional transportation agencies develop freight-oriented programs and policies for winter road maintenance programs, structural and geometric pavement design, highway life cycle analysis, long-range transportation planning, and resiliency metrics and planning. For long-range planning, the model developed in this paper can be incorporated into climate change scenarios that predict increased occurrence of rainfall and snow. Predictions of delay associated with temporal and spatial shifts of truck traffic due to climate change scenarios would allow finer estimation of cost/benefit ratios for project prioritization. For the trucking industry, carriers need to understand how adverse weather conditions affect the spatial and temporal traffic patterns of the truck population (not just their own fleet) to better plan routes and schedules for their own drivers. More accurate estimates of routes, travel times, and mileage stemming from a better understanding of what affects those estimates helps to improve cost efficiency, specifically in the calculation of revenue and billable miles and estimated times of arrival (ETA).

This paper is organized as follows. The Literature review section summarizes the most related previous studies to this study. The Methodology section details the traffic and weather data sources and model specification. The Results section compares the ordinary least square (OLS) and dynamic SAR models. The paper concludes by highlighting significant findings, noting limitations, and suggesting future improvements.

4.3 Literature review

Since the body of work related to predicting the effects of weather on truck traffic volumes is considerably limited, this section presents a review of studies that examined weather

effects on all types of vehicles. The review is separated into two sections: Insights into Weather Effects and Prior Model Specifications.

4.3.1 Insights into Weather Effects

Studies related to weather effects of winter storms on total traffic date back to the early 1990's (e.g., see Hanbali & Kuemmel, 1993; Knapp & Smithson, 2000; Maze, Crum, & Burchett, 2005; Maze, Agarwal, & Burchett, 2006; Datla & Sharma, 2010; Cools, Moons, Creemers, & Wets, 2010). Overall, these studies showed statistically significant reductions in total traffic volumes resulting from winter storm events. Hanbali and Kuemmel (1993) conducted a regional analysis covering 11 sites across New York, Wisconsin, Minnesota, and Illinois and found reductions in total traffic volume between 8% and 56%, depending on the depth of the snowfall. In Iowa, Knapp and Smithson (2000) reported a reduction in total traffic volume between 16% and 47% during winter storm events characterized by more than four-hour durations of snowfall at 0.51 cm (0.2 inches) per hour. Knapp and Smithson (2000) showed that snowfall intensity and total snowfall could be used to predict the percent reduction in total traffic volume. Maze, Crum, and Burchett (2005) found that strong wind and reduced visibility due to snow led to traffic volume reductions as great as 80%. Datla and Sharma (2010) reported reductions of around 30% during periods with air temperatures below -25°C and reductions of 51% during periods of snowfall of 30 cm (12 inches) or more in Alberta, Canada. Datla and Sharma (2010) found that a reduction in traffic volume due to snow and cold varies with day of week, hour of day, type of highway, and intensity of cold with traffic volume reductions of 80% during snow storms when the visibility was less than a quarter mile and wind speed was more than 40 mph. Moreover, they were able to show that roads carrying non-discretionary trips experienced less volume reduction (0.5% - 1.7%) than the roads that carry recreational trips

(0.5% - 3.15%) when considering historical traffic data, snow depth, and temperature (Datla & Sharma, 2010).

While the above mentioned studies focused on adverse winter storm effects, weather conditions like temperature, rainfall, wind speed, etc. had also been shown to affect traffic volumes (Keay & Simmonds, 2005; Datla & Sharma, 2010; Cools, Moons, Creemers, & Wets, 2010; Fu, Lam, & Meng, 2014; Liu, Li, Li, & Shang, 2015). Keay and Simmonds (2005) examined the effect of rainfall on total traffic, developing two models, one for daytime and another for nighttime conditions, using historical traffic volumes, day of week, and rainfall as independent variables. They found that 2 mm to 5 mm rainfall in the spring reduced traffic volume by 3.43%. Liu, Li, Li, and Shang (2015) calculated the percentage change rate of traffic volume due to rainfall finding that traffic volume decreased by 6% to 14% depending on the intensity of rainfall. Fu, Lam, and Meng (2014) showed that frequent rainfall significantly affected daily activity travel patterns in multi-modal transit network. Cools, Moons, Creemers, and Wets (2010) found that the changes in travel behavior in response to these weather conditions were highly dependent on trip purpose.

Few studies modeled the effects of weather on truck traffic separately from that of total traffic due in part to limited availability of truck count data (Roh, Datla, & Sharma, 2013; Roh, Sharma, Sahu, & Datla, 2015; Bardal, 2017). Models to explain truck volume changes separately from passenger traffic are necessary to capture spatial and temporal variations in truck traffic volumes that are not observed for passenger traffic. Compared to passenger vehicles, trucks were less likely to cancel trips due to inclement weather conditions (Datla, Sahu, Roh, & Sharma, 2013; Maze, Crum, & Burchett, 2005). Roh, Datla, and Sharma, (2013) developed models predicting passenger car and freight truck volume based on snowfall, temperature, a snowfall-

temperature interaction term, and a four-year average of daily truck volume for a given day of the week and day of the year. Roh, Sharma, Sahu, and Datla, (2015) found that truck traffic increased during winter storms, possibly due to trucks shifting away from secondary highways to primary highways that had higher priority in winter maintenance programs and that the effect was similar for weekends and weekdays. Bardal (2017) found that the adverse weather conditions reduced traffic volume, particularly to passenger traffic, and that temperature had a small but significant effect on truck traffic volume. The study also showed that the volume reduction was relatively low.

Models of the effects of weather on truck traffic volume are limited, in part, due to the sparsity of static traffic sensors that distinguish passenger vehicles from trucks. To overcome such limitations, researchers have started to use historical truck Global Positioning System (GPS) data in lieu of static traffic sensor data. Pierce and Short (2012) used truck GPS data to show spatial volume shifts in truck traffic caused by flooding along Arkansas Interstate 40 in May 2011. The historical data revealed that many trucks chose regional detours to circumvent the flooding closure (Pierce & Short, 2012). However, this study was observational and did not develop predictive models to relate weather conditions to truck rerouting patterns.

In summary, while previous studies modeled the direct effects of adverse weather on total traffic volumes, very few studies considered the particular implications for trucks separately. Moreover, the existing studies did not capture the rerouting behavior of trucks as they relied on simple linear regression models that cannot show spatial and temporal correlations. Hence, a more advanced modeling technique like spatiotemporal model should be used to capture such effects.

4.3.2 Prior Model Specifications

Previous studies used different methods to understand traffic volume variations due to weather events including hypothesis testing (Cools, Moons, Creemers, & Wets, 2010), rule-based algorithms (Fu, Lam, & Meng, 2014), structural equation models (Bardal, 2017), and ordinary least square (OLS) regression models (Kockelman, 1998; Knapp & Smithson, 2000; Keay & Simmonds, 2005; Datla & Sharma, 2010; Roh, Datla, & Sharma, 2013; Roh, Sharma, & Datla, 2014; Dong, Xiong, Shao, & Zhang, 2015; Liu, Li, Li, & Shang, 2015; Tessier, Morency, & Saunier, 2016; Rowell et al., 2012; Hanbali & Kuemmel, 1993; Knapp & Smithson, 2000). Though OLS models can explain a normally distributed linear relationship, they are not suitable when dependent or independent variables show spatial and temporal autocorrelation. When spatial autocorrelation is suspected, spatial regression techniques are more appropriate than OLS, because OLS estimators are biased and inconsistent in the presence of spatial autocorrelation (LeSage & Pace, 2009).

As indicated by previous studies, trucks are more likely to reroute rather than opt out of traveling (Winston & Shirley, 2004). This means that truck volumes on the link experiencing adverse weather are likely to be affected and that neighboring links (along the detour) may also be affected. To account for spatial and temporal autocorrelation, Dong, Xiong, Shao, and Zhang (2015) used a spatial-temporal model for predicting freeway network total traffic flow. They stated that temporal factors could predict traffic flow on a congestion-free network while spatial factors could predict flow-drop during congestion. They also showed that a spatial-temporal model could predict traffic flow more accurately, since the average prediction accuracy of the model with spatial considerations was 9% higher than a linear regression model. Although the

authors accounted for spatial and temporal autocorrelation, the model did not attempt to separate the unique effects of weather on truck traffic to predict rerouting behaviors.

4.4 Methods

This study applied a dynamic spatial panel regression technique that relates variations in truck traffic patterns to weather conditions. There were several reasons to consider spatiotemporal autocorrelation in truck traffic volumes as they related to weather conditions. First, consider a fixed volume of truck traffic between an origin and destination (OD). In the event of adverse weather affecting the primary route between the OD pair, truck traffic will shift to an alternate route rather than cancel the trip (Datla, Sahu, Roh, & Sharma, 2013). This means a spatial autocorrelation may exist in traffic volumes such that low volumes along the main route due to adverse weather correspond to higher volumes along neighboring alternate routes. With the strategic placement of point sensors in a network, i.e. along primary and alternate routes, detection of rerouting may be possible (Hyun, 2016). Second, due to the inherent form of the highway network, spatial patterns of dependent and independent variables may exhibit spatial non-stationarity. For instance, the density of the road network differs across each region. In regions with high network density, detours around adverse weather may be more feasible compared to regions of low network density (CPCS, 2018). Thus, there may be spatial correlation in traffic volumes if network density is not explicitly captured as an independent variable. Lastly, willingness to delay a trip due to a weather event may be contingent on the commodity transported, e.g., refrigerated and perishable goods would be more sensitive to delays than would manufactured products (Winston & Shirley, 2004). As freight trip generation is tied to regional land uses and seasonality (FHWA, 2017), it is possible that spatiotemporal

autocorrelation exists due to the movement of specific commodities within a region or along a particular highway route at any particular time.

Hence, a dynamic spatial autoregressive panel model with spatial fixed effects to measure the spatiotemporal effects of weather variables on daily truck traffic was developed in this paper. This uniquely corrected for the limitation of previous studies which neglected to consider spatial and temporal autocorrelation that occurred when analyzing the effects of weather on truck traffic patterns. After detailing the data sources and pre-processing steps, a discussion of the model specification is presented in this section.

4.4.1 Data Collection and Pre-Processing

Two types of data were used in this study: (i) traffic volume data by vehicle class, and (ii) weather data. Similar to previous studies, this study used traffic data from fixed sensors such as Automatic Traffic Recorders (ATR) to obtain total traffic volumes and Weigh-in-Motion (WIM) sensors to obtain truck traffic volumes. Of all traffic sensors types, WIM provide the highest level of detail about the vehicle population. WIM sensors measure axle configuration, axle weight, vehicle length, and speed to predict vehicle type according to the commonly referenced Federal Highway Administration (FHWA) Scheme F (FHWA, 2013). Traffic volume data by vehicle class was gathered from 18 WIM stations located in Arkansas (Figure 4.1) and distinguished by the road functional class of the location. Information on truck type allowed analysis of weather-related impacts to be determined solely for the truck population. In this study, hourly volumes of trucks in FHWA classes five through 13 were used to calculate daily truck volume. Classes five through 13 correspond to common freight carrying trucks (vehicles in classes 1 through four are passenger vehicles or light duty trucks not carrying freight). All holidays were removed from the data before calculating model parameters.

Since WIM stations in Arkansas do not monitor weather, it is necessary to gather weather data from an alternate source. Daily weather data were obtained from MERRA through the LTPP InfoPave Climate Tool (FHWA, 2018b). MERRA climate data includes average humidity, surface runoff, snowfall, etc., weather parameters. Average humidity represents the probability of rain for a day while surface runoff measures the flow of water on surface due to rain. Each MERRA zone provides the daily average weather condition for an area of 1,225 (35×35 mi) square miles. Although the weather data used in this paper represents average conditions over an area (e.g., not at the specific WIM site location), it was assumed that this resolution of weather data was approximate to what truck drivers would have access to when making routing or other travel decisions.

Since there were approximately 45 MERRA zones in Arkansas, each MERRA zone was assumed to capture the weather conditions at each WIM site. Previous studies suggested that homogeneous weather patterns were found within a radius of 10-16 miles around a weather station (Roh, Datla, & Sharma, 2013). However, correlation analysis of weather variables and distance between WIM and land weather stations of the National Oceanic and Atmospheric Administration (NOAA) conducted in this study found that, for Arkansas, a radius of 65 miles around the weather station was appropriate given the more homogeneous weather patterns in Arkansas. By assigning the closest MERRA zone to each WIM site, a maximum radius of 25 miles from the WIM station to the MERRA zone centroid could be achieved and was within the bounds defined by the weather-distance correlation analysis for Arkansas (Figure 4.1). Note that the MERRA weather data was used in favor of the weather data from the National Oceanic and Atmospheric Administration (NOAA) because the NOAA weather data was not temporally continuous during the study period.

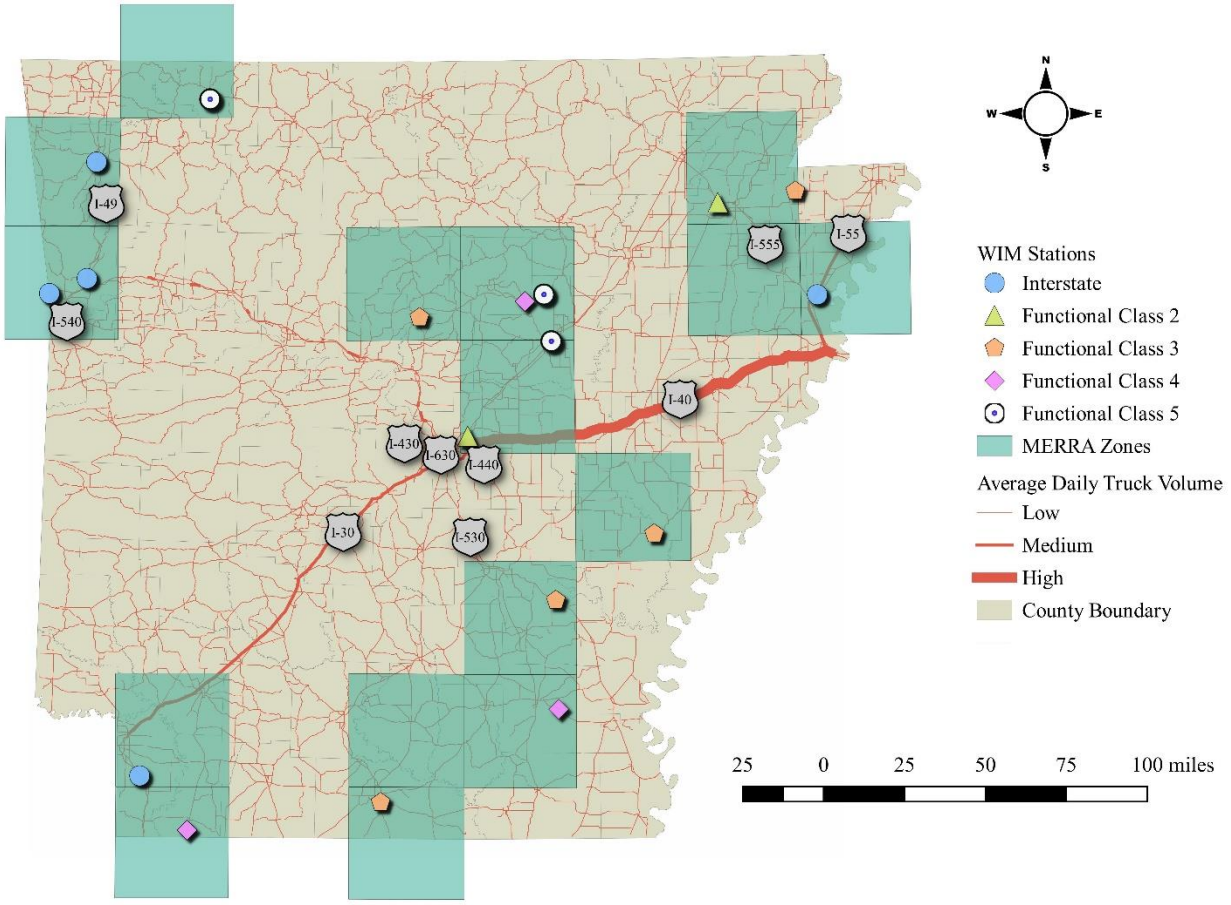


Figure 4.1 WIM stations and MERRA weather zones included in the study

4.4.2 Variable Specification

The percentage change in daily truck volume ($y_{d,s,yr}$) (Eq. 4.1) compared to the AADTT ($AADTT_{s,yr}$) (Eq. 4.2) of each WIM station (s) served as the dependent variable in the model.

$$y_{d,s,yr} = \frac{v_{d,s,yr} - AADTT_{s,yr}}{AADTT_{s,yr}} \times 100\ \% \quad (4.1)$$

$$AADTT_{s,yr} = \frac{\sum_{d=0}^{N} v_{d,s,yr}}{N} \quad (4.2)$$

Where,

$y_{d,s,yr}$ = Percentage change in daily truck volume to the AADTT for a particular date *d* of station *s* in year *yr*

$v_{d,s,yr}$ = Truck volume for a particular date *d* of station *s* in year *yr*

$AADTT_{s,\,yr}$ = Annual Average Daily Truck Traffic of WIM station *s* in year *yr*

N = Number of days in year *yr* for which data was recorded at the WIM site

The Expected Daily Volume Factor (EDVF) captured the historical trend in truck volume (Roh et al., 2013). It was calculated by the average proportion of daily truck volume compared to the Annual Average Daily Truck Traffic (AADTT) (Eq. 4.3) over a five-year period (2011 to 2015). This period was chosen based on available WIM and MERRA data. An EDVF value greater than one (> 1) indicates a historical higher daily truck volume and EDVF value lower than one (<1) indicates a historical lower daily truck volume compared to the AADTT.

$$EDVF_{i,j,k,s} = \frac{\sum_{r=2015}^{r=2011}\left(DVF_{i,j,k,yr,s}\right)}{5}\ ,\forall i,j,k\ \&\ s \qquad (4.3)$$

Where,

$EDVF_{i,j,k,s}$ = Expected daily volume factor for a particular day i of the week (e.g., Monday, Tuesday), a particular week j of the month (i.e. Week 1 – Week 5), a particular month k of the year (i.e. January – December) of station s

$DVF_{i,j,k,yr,s}$ = Daily Volume Factor for a particular day i of the week (e.g., Monday, Tuesday), a particular week j of the month (i.e. Week 1 – Week 5), a particular month k of the year (i.e. January – December) of station s in year yr (i.e. 2011-2015) calculated as

$$DVF_{i,j,k,yr,s} = \frac{\bar{v}_{i,j,k,yr,s}}{AADTT_{s,yr}} \qquad (4.4)$$

Where,

$AADTT_{s,\,yr}$ = Annual Average Daily Truck Traffic of WIM station *s* in year y*r*

$\bar{v}_{i,j,k,yr,s}$ = Average truck volume for a particular day *i* of the week (e.g., Monday, Tuesday), a particular week *j* of the month (i.e. Week 1 – Week 5), a particular month *k* of the year (i.e. January – December) of station *s* in year *yr* (i.e. 2010-2015)

In addition to EDVF, a total of eight independent variables were specified in the model, covering three categories: (a) weather variables, (b) historical traffic volume (e.g., EDVF), and (c) temporal variables.

Eleven weather variables were collected from MERRA for this study (Table 4.1). An analysis of multi-collinearity showed that precipitation, evaporation, infiltration, and runoff were highly correlated; average temperature, maximum temperature, and minimum temperature were correlated; and snow mass and snowfall were correlated. Based on the multi-collinearity analysis, one weather-related variable was selected from each weather "category", i.e. rain, temperature, snow, and a backward stepwise elimination method was used to estimate an appropriate regression equation consisting of three weather variables, i.e. average humidity, surface runoff, and snow mass. Multi-collinearity was within the acceptable range for these variables (the largest variance inflation factor was <4). Descriptive statistics of the independent variables (Table 4.2) showed that weather variables were continuous and changed over time (between) and space (within). Note that Table 4.2 includes only the overall variation for temporal variables representing season and day of week.

Table 4.1 MERRA Weather Variables

Weather Variables	Definition	Units
Average Humidity	Average hourly relative humidity for the day.	%
Precipitation	Water equivalent of total surface precipitation over time (day).	mm
Evaporation	Surface evaporation over time (day).	mm
Infiltration	Water on the ground surface enters the soil over time (day).	mm
Surface Runoff	Water flow due to rain over the Earth's surface for a day.	mm
Snow mass	Snow mass over an area.	kgm^{-2}
Snowfall	Depth of snowfall.	mm
Maximum Wind Velocity	Maximum hourly average wind velocity 2 meters above MERRA centroid elevation for the day.	ms^{-1}
Average Temperature	Average of the hourly air temperatures 2 meters above the MERRA centroid.	˚C
Maximum Temperature	Maximum hourly air temperature 2 meters above elevation of MERRA cell centroid.	˚C
Minimum Temperature	Minimum hourly air temperature 2 meters above elevation of MERRA cell centroid.	˚C

Table 4.2 Independent Variables Included in Models

Independent Variables			Mean	Std. Dev.	Min	Max
Weather Variables	Avg. Humidity (%)	Overall	74.21	11.90	38.00	97.00
		Between		1.31	70.82	76.24
		Within		11.83	39.90	98.39
	Surface Runoff (mm)	Overall	0.68	3.56	0.00	73.90
		Between		0.41	0.18	1.39
		Within		3.54	0.00	73.19
	Snow Mass (kgm^{-2})	Overall	1.81	12.12	0.00	179.20
		Between		1.25	0.13	4.04
		Within		12.06	0.00	176.98
Historical Traffic Volumes	Expected Daily Volume Factor (EDVF)	Overall	0.98	0.32	0.25	5.15
		Between		0.02	0.92	1.02
		Within		0.32	0.26	5.21
Temporal Variables	Weekend (Saturday, Sunday)	Overall	0.32	0.46	0.00	1.00
	Fall (September, October, November)	Overall	0.39	0.49	0.00	1.00
	Winter (December, January, February)	Overall	0.29	0.45	0.00	1.00
	Spring (March, April, May)	Overall	0.24	0.43	0.00	1.00
	Summer (June, July, August)	Overall	0.08	0.27	0.00	1.00

Observations: Overall, N = 2052 records; Between, T = 114 days; Within, n = 18 stations

4.4.3 Model Specification

A balanced panel dataset, e.g., multi-dimensional data involving measurements over time, consisting of n spatial units (n= 18 stations) observed for T periods (T=114 days) was used in this study. Panel data increases the efficiency of model estimation and captures more complicated behavioral hypotheses, including effects (Elhorst, 2013; Hsiao, 2005).

Since Ordinary Least Square (OLS) regression analysis was commonly used to explain the relationships among the weather variables and traffic volumes (Kockelman, 1998; Knapp & Smithson, 2000; Keay & Simmonds, 2005; Datla & Sharma, 2010; Roh, Datla, & Sharma, 2013; Roh, Sharma, & Datla, 2014), a non-spatial linear regression model was developed for comparison purposes and to facilitate selection of an appropriate spatial model. A pooled OLS model with special specific effects, but without spatial interaction effects for a panel data can be written as (Elhorst, 2014):

$$\boldsymbol{Y}_{it} = \alpha + \boldsymbol{\beta X}_{it} + u_i + \varepsilon_{it} \qquad (4.5)$$

Where,

i = an index for the cross-sectional dimension (stations)

t = an index for the time dimensions (days)

$\boldsymbol{Y_{it}}$ = percentage change in daily truck volume of station i for day t

$\boldsymbol{X_{it}}$ = a vector of explanatory variables (i.e. humidity) of station i for day t

$\boldsymbol{\beta}$ = the coefficient of explanatory variables $\boldsymbol{X_{it}}$

a = the coefficient of intercept

u_i = a spatial specific effect;

The standard reasoning behind spatial specific effects is that they control for all space-specific time-invariant variables whose omission could bias the estimates in

a typical cross-sectional study. It is assumed that $\mu \sim N(0, \sigma_u^2)$ in the random-effects case, while the μ is a vector of parameters to be estimated in the fixed-effects variant.

ε_{it} = is an independently and identically distributed error term for station *i* in day *t* with zero mean and variance σ^2

The parameters defined from the dataset used in this study exhibited 'fixed effects' in the parameter distributions. A dummy variable was introduced for each time period. While pooled OLS or fixed-effects Generalized Least Squares (GLS) are commonly used to predict the traffic volumes using panel data, OLS or GLS estimates may be biased and inconsistent in the presence of spatial effects (LeSage & Pace, 2009). Instead, a spatial regression model is required. Spatial regression models explain the effects of the independent variables after removing the effects of spatial autocorrelation. Based on Moran's I statistic (Cliff, Ord, Haggett, & Versey, 1981), spatial interactions were indeed present within the dataset (p-value<0.01).

Specification of a spatiotemporal model is based on the type of spatial interaction effects among the error terms, i.e. endogenous, exogenous, and interaction effects. Endogenous effects explain that the value of a dependent variable y at location A depends on the change in the neighboring dependent variable y at location B (Figure 4.2). Exogenous effects explain that the value of a dependent variable y at location A depends on the change in an independent variable x at the neighboring location B (Figure 4.2). Interaction effects among the error terms explain that the omitted determinants of the dependent variable are spatially auto-correlated (Figure 4.2) (Elhorst, 2013).

Dependent variable *y* at ***A*** ⇿ Dependent variable *y* at ***B***

Independent variable *x* at ***B*** ⇿ Dependent variable *y* at ***A***

Figure 4.2 Spatial interaction effects

There are two common types of spatial regression models: Spatial Error models and Spatial Autoregressive models (SAR). Spatial Error models are appropriate when errors are spatially correlated due to random features associated with location and when both the dependent and the independent variables have spatial autocorrelation. Spatial Error models the effect of the independent variables on the dependent variable after removing the effect of spatial dependencies from dependent and independent variables (Eq. 4.6) (Belotti, Hughes, & Mortari, 2017).

$$y_{it} = X_{it}\beta + \mu_i + \phi_{it}$$
$$\phi_{it} = \lambda W_{ij}\, \phi_{jt} + \varepsilon_{it} \qquad (4.6)$$

Where,

Φ_{it} = reflects the spatially auto-correlated error term

λ = spatial autoregressive parameter

W_{ij} = an element of a spatial weight matrix W describing the spatial arrangement of the units in the sample. It is assumed that W is a pre-specified non-negative matrix of order N.

Other terms as previously defined

SAR models are appropriate when the dependent variable is spatially correlated meaning that spatial dependencies exist directly among the levels of the dependent variable. SAR residuals show a random pattern while the OLS residuals have a non-random pattern and exhibit clustering. SAR models the effect of the independent variables on the dependent variable after

removing the effect of spatial dependencies from the dependent variable (Eq. 4.7) (Belotti, Hughes, & Mortari, 2017).

$$y_{it} = \rho W_{ij} y_{jt} + X_{it}\beta + u_i + \varepsilon_{it} \quad (4.7)$$

Where,

ρ = spatial autoregressive parameter

Other terms as previously defined

Following Anselin (2005), a Lagrange Multipliers (LM) test was used to determine the specific spatial dependence of the data. The LM test showed significant values (LM statistic: 4.90, significant at 95% level of confidence) only for the SAR model indicating the SAR model was more appropriate than the Spatial Error model for the type of spatial dependency in the data. Therefore, a SAR model was applied to predict the effect of weather events on daily truck volumes.

Since the dependent variable of this study was both space and time lagged, a dynamic linear spatial dependence model, specifically a dynamic SAR model, was used (Eq. 4.8) (Debarsy, Ertur, & LeSage, 2012; Elhorst, 2013).

$$y_{it} = \tau y_{it-1} + \rho W y_{it} + \eta W y_{it-1} + X_{it}\beta + u_i + \varepsilon_{it} \quad (4.8)$$

Where,

τ = time dependence autoregressive parameter

η = spatiotemporal diffusion parameter

Other terms as previously defined

a. Spatial Weight Matrix

Spatial models depend on the spatial weight matrix. The spatial-weight matrix implemented in this study followed from Tobler's first law of geography- "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Thus, W was created based on the inverse distance matrix. *Following* Anselin (2005), *W had* a minimum distance (*dmin*) to ensure that every WIM station had at least one neighbor station. The spatial effect of one station on another station decreased, when the distance between them increased. This study used an 18 × 18 spatial-weight matrix, W in this study. W was calculated using the longitudes and latitudes of the 18 WIM stations. Each element *wij* of *W* was defined as:

$w_{ij} = 1/d_{ij}$ where $d_{ij} \geq d_{min}$, $i, j = 1, \ldots, N$, $i \neq j$ and

d_{ij} = the distance between the centroids of WIM station *i* and WIM station *j*

$\overline{\boldsymbol{W}}$ is the row normalized form of weight matrix $\boldsymbol{W}$, where

$\overline{w_{ij}} = \frac{w_{ij}}{\sum_{j=1}^{N} w_{ij}}$; by convention, $w_{ij} = \overline{w_{ij}} = 0$ for $i = j$ due to the exclusion of self-station.

b. Interpreting Results

Interpretation of a dynamic SAR model differs from that of conventional OLS, because of the space and time lag terms, which create feedback effects between neighboring stations over time. A dynamic SAR model explains the effect of a change in an independent variables (historical traffic volume, weather, and season) for a specific station on the dependent variable (percentage change in daily truck volume) at station itself (direct impact) and, potentially, on all other stations (indirect impacts) both on the same day (short-term) and previous and past days (long-term). This implies the existence of direct, indirect, and total marginal impacts for both short-term and long-term periods (LeSage & Pace, 2009; Elhorst, 2013). Direct impacts

represent the average impact on each station of changes in the explanatory variables for the station itself, including the feedback passing through neighboring stations and back to each station. The indirect impact represents the impacts on other stations only, also known as spatial spillovers. The total impact is the sum of direct and indirect impacts (Belotti, Hughes, & Mortari, 2017). This model also measures how region A's dependent variable responds over time to changes in a given time period. This is also referred as diffusion effects. The short-term (Eq. 4.9) and long-term (Eq. 4.10) impacts quantify the effect of explanatory variables at time t on the dependent variables of each region at various time horizons, t +T (Debarsy, Ertur, & LeSage, 2012; Elhorst, 2013).

$$[(I - \rho W)^{-1} \times (\beta_k I_N + \theta_k W)] \quad (4.9)$$

$$[(1 - \tau)I - (\rho + \eta)W)^{-1} \times (\beta_k I_N + \theta_k W)] \quad (4.10)$$

Where,

I = the N × N identity matrix

Other terms as previously defined

4.5 Results

A dynamic SAR model was estimated to determine the effect of weather events on truck traffic volume. A Quasi-Maximum likelihood estimation for the fixed effects dynamic SAR model was carried out using statistical software, i.e. STATA 14. Table 4.3 compares the OLS and dynamic SAR models showing only the total short and long term effects for the dynamic SAR model. The direct and indirect effects (which sum to the total effects) for the dynamic SAR model are shown in Table 4.4.

The spatial autoregressive parameter rho (ρ) was positive (0.37) and statistically significant, which reflected the spatial dependence inherent in the data (Belotti, Hughes, &

Mortari, 2017). Recall that the SAR model explains spatial autocorrelation in the dependent variable, e.g., the percentage change in daily truck volume ($y_{d,s,yr}$).

The coefficient value of average humidity in the SAR model explained that if average humidity near a WIM station increased by 10 percent, the daily truck volume of that station decreased by 2.8 percent within a short-term period and 3.1 percent after a long-term period (Table 4.3). Moreover, if there were 2.2 pounds of snow per square foot (10 kgm-2) accumulated, the daily truck volume decreased by 2.3 percent within a short-term period and 2.6 percent after a long-term period (Table 4.3). Previous studies found that the 8.54 inches snowfall that accumulates to 2.2 pounds of snow per square foot reduced passenger car volume by 56 percent (Roh, Sharma, & Datla, 2014). Alternatively, if there were 4 inches (100 mm) of runoff due to rain, daily truck volume decreased by 70 percent and 77 percent after a short and long-term periods, respectively (Table 4.3).

The result also showed that historical volume, EDVF, had a significant positive effect on the percent change in daily truck volume ($y_{d,s,yr}$) for both OLS and dynamic SAR models. The spatial model predicted that if EDVF value increased by one unit, the daily truck volume increased by 19.11%. Alternatively, temporal variables, i.e. weekend and season, had significant effects on daily truck volume. The coefficient value of the spatial (main) model showed that daily truck volume decreased by 23.09% on a weekend (i.e. Saturday and Sunday) compared to a weekday.

The dependent variable of this study was both time-lagged and space-time-lagged. Hence, the dynamic SAR model measured the effect of a time-dependence autoregressive parameter (τ) and a spatiotemporal diffusion parameter (η) on the dependent variable (Table 4.3). The statistically significant positive effect of the time-dependence autoregressive parameter (τ)

explained that if truck volume at a station decreased by one percent on a specific day t, it decreased by 0.55 percent at the same station on the next day t+1. Alternatively, the statistically significant negative effect of spatiotemporal diffusion parameter (η) explained that if truck volume of a station decreased by one percent on a specific day t, it increased by 0.49 percent at the neighboring stations on the next day t+1, and thus captured the rerouting behavior. The results showed that both direct and indirect effects of weather variables were negative and significant in the short-term (Table 4.4). The shift in truck traffic from the main to an alternate route due to adverse weather did not happen instantaneously, but after some delay. Hence, the long-term indirect effects of adverse weather variables were positive, while the long-term direct effects were negative. This key finding effectively captured the rerouting behavior of trucks as they shift to alternate routes in response to adverse weather in a region. Assuming fixed OD demand flows, truck drivers already on the route impacted by adverse weather cannot alter their routes, and hence truck volume on the impacted route does not change immediately. After some delay, a day or more, truck drivers shift to alternate routes such that increases in truck volumes on neighboring routes are observed. For instance, if a road experienced snow mass accumulation of approximately two pounds per square foot, a three percent truck volume decreased over a one-day time horizon for that road as a result of truck drivers rerouting, i.e., the estimated long-term direct effect (Table 4.4). Concurrently, neighboring roads experienced an almost one percent increase in truck volume over the one-day time horizon, i.e., the estimated long-term indirect effect (Table 4.4).

In summary, the dynamic SAR model captured the short-term and the long-term effects (Table 4.3) with direct and indirect impacts of the weather variables (Table 4.4). Alternatively, OLS did not capture these effects. In addition, the higher R^2 value of the dynamic SAR model

and lower Akaike information criterion (AIC) value showed the dynamic SAR to be a better fit than the OLS model (Table 4.3).

Table 4.3 Results of OLS Model and Dynamic SAR Model

Independent Variables	OLS Regression	Dynamic SAR with Spatial Fixed-Effects		
		Main	Short-Term Total	Long-Term Total
Avg. Humidity	-0.31***	-0.17***	-0.28***	-0.31***
Surface Runoff	-0.42***	-0.48**	-0.70**	-0.77**
Snow Mass	-0.29***	-0.15***	-0.23***	-0.26***
EDVF	28.06***	19.11**	30.43**	33.66**
Weekend	-38.25***	-23.09***	-36.75***	-40.73***
Base: Winter				
Fall	-7.93***	-4.43**	-7.16**	-7.93**
Summer	-2.21	-1.09	-1.64	-1.80
Spring	-5.16***	-2.50	-3.98	-4.39
Constant	12.86**			
Time-Dependence (τ)		0.55***		
Spatiotemporal (η)		-0.49***		
Spatial, rho *(ρ)*		0.37***		
R-squared:				
Within	0.52	0.67		
Between	0.15	0.90		
Overall	0.52	0.67		
AIC	18965.96	17996.11		
BIC	19016.60	18063.52		

***significant at 99% confidence level;
**significant at 95% confidence level;
*significant at 90% confidence level

Table 4.4 Direct and Indirect Impact of Dynamic SAR Model with Spatial Fixed Effects

Independent Variables	Short-Term Impact		Long-Term Impact	
	Direct	Indirect	Direct	Indirect
Avg. Humidity	-0.18***	-0.10***	-0.40***	0.09**
Surface Runoff	-0.45*	-0.24**	-1.00*	0.23
Snow Mass	-0.15***	-0.08***	-0.33***	0.07***

4.6 Discussion

Expanding on prior work, this study showed, through a dynamic SAR model, that not only does weather impact truck volume but there are distinct and significant effects in both the spatial and temporal changes in truck volume.

The model explained that snow mass had a significant spillover (indirect) effect at the 99% level of confidence in both the short and long-term. Snow mass, unlike snowfall, captures the effect of snow accumulation that creates obstacles for the movement of large trucks over longer periods. Hence, the long-term direct effect (-0.33) of snow mass was approximately twice as high as the short-term direct effect (-0.15). Alternatively, the spillover effect showed that truck volume started increasing (+0.07) at neighboring stations after a long-term period. Snow requires special road winter management e.g., snow removal, deicing salt, etc. which may take time and cause disruptions to traffic that extend for longer durations. The impacts of snow mass were observed in the percent change in daily truck volume (Figure 4.3).

The estimated parameters of the dynamic SAR model also showed negative and significant effects on truck volume due to increased average humidity. Average humidity is an indicator of fog and encapsulates the effects of daily temperature and dew point temperature (NOAA, 2015). Like snow mass, it also had higher direct effect over the long-term (-0.40) than the short-term (-0.18). Though fog, as captured by average humidity, does not result in road obstacles or closures like snow mass, it affects visibility leading to unsafe driving conditions and travel delays. As observed in the model, the truck drivers chose to reroute to neighboring roads increasing daily truck volume by 0.09 percent.

Interestingly, the observed impacts of surface runoff differed from those of humidity and snow mass. The study found that surface runoff had the highest negative direct effect for both

short-term (-0.45) and long-term (-1.00). The long-term effect of surface runoff showed that if there was 0.4 inch (10 mm) of runoff on a particular road link, daily truck volume at that station decreased by 10 percent over the long-term. However, though surface runoff had the highest negative direct effect, it did not have any long-term spillover effects. This could be due to runoff being an immediate impediment that was accommodated by drivers by delaying trips for only a couple hours rather than causing drivers to shift routes. Contrast that to snow mass which created a longer term driving impediment, e.g., days, and thus drivers chose to reroute. Surface runoff is an indicator of flash floods and encapsulates the effects of heavy rainfall from storm events.

The results showed that weekdays (Monday, Tuesday, Wednesday, Thursday, and Friday) had higher daily truck volume compared to weekends (Saturday and Sunday) in line with previous studies, e.g., Hallenbeck, Rice, Smith, Cornell-Martinez, and Wilkinson (1997) showed that Monday truck volumes tended to be the lowest while Wednesday had the highest truck volume of a week (excluding weekends). Seasonally, lower truck volume was seen in the winter months while higher truck volume was observed in the late spring through early fall (Hallenbeck, Rice, Smith, Cornell-Martinez, & Wilkinson, 1997). However, this study found that Arkansas experienced lower truck volumes in fall than in the winter. Higher daily truck volumes in Arkansas may be due to the movement of agricultural goods after the fall harvest season. Considering the dominance of agricultural industries in Arkansas this is a feasible conclusion.

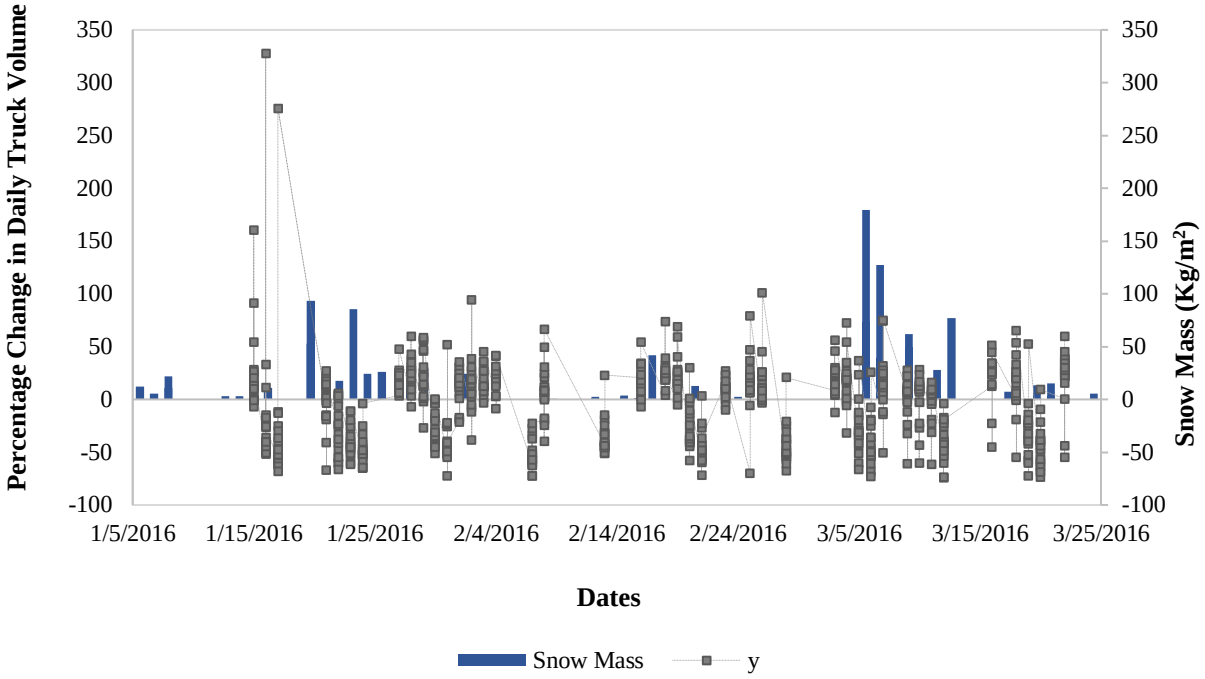


Figure 4.3 Example of the effects of snow mass accumulation on daily truck volumes

4.7 Conclusion

This study investigated the spatial and temporal effects of adverse weather conditions on daily truck traffic volume through the application of a dynamic Spatial Autoregressive (SAR) panel model with spatial fixed effects that captured the rerouting behavior. The estimated model explained how one unit change in weather related variables (i.e., snow mass, humidity, and surface runoff etc.) could affect daily truck traffic volume of a route and its neighboring routes relative to the AADTT, controlling for day of week and season. The paper used a historical truck volume, e.g., Expected Daily Volume Factor or EDVF, computed over a five-year period (2011-2015) to predict future truck volume (2016).

The results showed that changes in truck traffic volume due to weather conditions exhibited spatial (direct and indirect) and temporal shifts (short and long term effects) that resulted in rerouting. Among three weather variables, surface runoff caused the highest volume

reduction, 0.70 percent and 0.77 percent for short-term and long-term, respectively. For long-term effects, snow mass caused 0.26 percent volume reduction while average humidity caused 0.31 percent volume reduction. The study showed that daily truck volume followed historical patterns, increasing by 30.43 percent, if EDVF increased by one unit after a short-term. The percentage change in daily truck volume also depended on the day of a week. Weekends (i.e. Sunday and Saturday) had comparatively lower (36.75 percent reduction) truck volume than weekdays (e.g., Tuesday, Wednesday, etc.) after a short-term. Additionally, truck volume was lower during the fall (September, October, and November) than in the winter (December, January, and February).

The estimated model showed that the spatial autoregressive parameter (ρ) was statistically significant indicating that truck volume had spatial dependency and should be analyzed with a spatial regression model, rather than a more standard OLS approach. A comparison between coefficients estimated via OLS and a dynamic SAR model illustrated the perils of using OLS in the presence of spatial autocorrelation. Moreover, the dynamic SAR model was able to capture temporal and spatial shifts in truck volumes. This is important considering the behavioral differences between passenger and freight travel decisions in light of weather events. Trucks follow rigid schedules for pickup and delivery and do not cancel trips (Winston & Shirley, 2004). Through a spatiotemporal model, this paper was able to capture rerouting behaviors through temporal and spatial shifts in truck volume.

The prediction of both spatial and temporal effects of weather on truck traffic volumes can support and improve long-range transportation planning as well as maintenance operations. For instance, the predictive model developed in this paper can help state Departments of Transportation (DOTs) or local transportation agencies prioritize road maintenance and

inclement weather operations for freight traffic. As an example, given a snow mass of approximately 2.2 pounds per sq. feet is predicted for a particular road segment with an average daily truck volume of 10,000 trucks, the dynamic SAR model could be used by decision makers to estimate that neighboring alternate roads will observe an increase of 70 trucks (0.7 percent) over the next several days. This could lead to decisions on where to apply deicing treatment along neighboring routes. Following the same example, but considering long-range planning contexts, the estimated number of rerouted trucks per day along with the estimated length of the detour can be used to calculate user costs to generate cost/benefit ratios needed for project prioritization. This would also be beneficial for resilience planning as a way to identify critical network links that may incur additional truck traffic during adverse weather conditions.

In the context of the trucking industry, delays caused by rerouting and re-scheduling that are not accounted for in the original route plan and schedule lead to cost inefficiencies. Consider the 70 trucks described in the previous example. If those drivers were to shift their routes, the additional mileage could exceed the billable mileage, lead to the need for additional required rest breaks, and delay the Estimated Time of Arrival (ETA). The model described in this paper can help shippers more accurately calculate billable miles by incorporating predictions of adverse weather conditions (Winston & Shirley, 2004).

While this study focused on the prediction of truck volume changes, it would be valuable further consider changes in Vehicle Miles Traveled (VMT) and Vehicle Hours Traveled (VHT), which is a combination of volume and route changes. Moreover, it would be equally valuable to include seasonal traffic variations within the historical traffic volume measure, AADT, which was used as an independent variable in the spatial regression model. AADT represents an average of daily and seasonal variability in traffic volumes but as an annual average does not

allow us to detect seasonal historical trends in traffic volumes. As used in our model, seasonal variability is captured by seasonality dummy variables but in future model specifications could also be captured through seasonal traffic volume measures at each site. In addition, a potential improvement of this model is to find the relationship between the types of cargo carried and rerouting due to weather conditions. A possible way to estimate VMT/VHT change and to consider cargo carried is to use anonymous truck Global Positioning System (GPS) data within a dynamic SAR model. Unlike static traffic data, e.g., Weigh-In-Motion (WIM) or AADT, which provide only point estimates of traffic volumes, GPS provides insights into route taken, trip length and duration, and origin-destination. Truck GPS data is spatially continuous and can be paired with weather data from MERRA to not only study changes in truck volume at each site, but to study the changes in VMT/VHT due to weather conditions. While existing studies used GPS data to look at historical changes in travel patterns due to weather events (Pierce & Short, 2012), a predictive model based approach like the one outlined in this paper would add to the understanding of the effects of weather and thus to the types of applications for such work. GPS data can also be used to correlate the type of cargo and the re-scheduling. Recent advances in distinguishing detailed truck characteristics from anonymous truck GPS data could be used to discriminate cargo types (Sun & Ban, 2013; Akter & Hernandez, 2019a; Akter & Hernandez, 2019b).

4.8 Acknowledgement

The authors thank the Southern Plains Transportation Center (SPTC) a University Transportation Center funded by the U.S. Department of Transportation, for sponsoring the project that lead to this paper.

4.9 Authors Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: S. Hernandez; data gathering and processing: T. Akter and K. Diaz; analysis and interpretation of results: T. Akter, S. Mitra, S. Hernandez; draft manuscript preparation: T. Akter. All authors reviewed the results and approved the final version of the manuscript.

4.10 References

Akter, T., & Hernandez, S. (2019a). Measuring the Effect of Weather Events on Long-Haul Truck Traffic Using Anonymous Truck GPS Data. Paper presented at the *AASHTO 2019 GIS for Transportation Symposium.*

Akter, T., & Hernandez, S. (2019b). Truck Activity Pattern Classification Using Anonymous Mobile Sensor Data. Paper presented at the *2019 Innovations in Freight Data Workshop.*

Anselin, L. (2005). Exploring Spatial Data with GeoDaTM: A Workbook. *Center for Spatially Integrated Social Science.*

Bardal, K. G. (2017). Impacts of Adverse Weather on Arctic Road Transport. *Journal of Transport Geography, 59,* 49-58. doi:10.1016/j.jtrangeo.2017.01.007.

Belotti, F., Hughes, G., & Mortari, A. P. (2017). Spatial Panel-Data Models Using Stata. *The Stata Journal, 17*(1), 139-180. doi:10.1177/1536867X1701700109.

Cliff, A. D., Ord, J. K., Haggett, P., & Versey, G. R. (1981). *Spatial Diffusion: An Historical Geography Of Epidemics In An Island Community* CUP Archive.

Commendatore, P., Kayam, S., & Kubin, I. (2015). *Complexity and Geographical Economics: Topics and Tools* (2015th ed.). Cham: Springer. doi:10.1007/978-3-319-12805-4 Retrieved from https://ebookcentral.proquest.com/lib/[SITE_ID]/detail.action?docID=1997943.

Cools, M., Moons, E., Creemers, L., & Wets, G. (2010). Changes in Travel Behavior in Response to Weather Conditions. *Transportation Research Record: Journal of the Transportation Research Board, 2157*(1), 22-28. doi:10.3141/2157-03.

CPCS. (2018). *COMPASS Freight Study.* Retrieved from http://www.compassidaho.org/documents/prodserv/CIM2040_20/FreightStudy2017_SystemReliability.pdf.

Datla, S., Sahu, P., Roh, H., & Sharma, S. (2013). A Comprehensive Analysis of the Association of Highway Traffic with Winter Weather Conditions. *Procedia - Social and Behavioral Sciences, 104,* 497-506. doi:10.1016/j.sbspro.2013.11.143.

Datla, S., & Sharma, S. (2010). Variation of Impact of Cold Temperature and Snowfall and Their Interaction on Traffic Volume. *Transportation Research Record: Journal of the Transportation Research Board, 2169*(1), 107-115. doi:10.3141/2169-12.

Debarsy, N., Ertur, C., & LeSage, J. (2012). Interpreting Dynamic Space-Time Panel Data Models. *Statistical Methodology*, Retrieved from https://pure.fundp.ac.be/portal/en/publications/interpreting-dynamic-spacetime-panel-data-models(14941fcc-8163-460f-96e5-f232825b9d31).html.

Dong, C., Xiong, Z., Shao, C., & Zhang, H. (2015). A Spatial-Temporal-Based State Space Approach for Freeway Network Traffic Flow Modelling and Prediction. *Transportmetrica A: Transport Science, 11*(7), 547-560. doi:10.1080/23249935.2015.1030003.

Elhorst, J. P. (2014). Spatial Panel Data Models. *Spatial econometrics* (pp. 37-93) Springer.

FHWA. (2013). Traffic Monitoring Guide. Retrieved from https://www.fhwa.dot.gov/policyinformation/tmguide/tmg_2013/vehicle-types.cfm.

FHWA. (2017). FHWA Freight Management and Operations - Freight and Congestion. Retrieved from https://ops.fhwa.dot.gov/freight/freight_analysis/freight_story/congestion.htm.

FHWA. (2018a). Status of The Nation's Highways, Bridges, and Transit Conditions and Performance: 23rd Edition: Part III: Highway Freight Transportation - Report to Congress. Retrieved from https://ops.fhwa.dot.gov/freight/infrastructure/nfn/rptc/cp23hwyfreight/iii_ch11.htm.

FHWA. (2018b). Long-Term Pavement Performance (LTPP). Retrieved from https://infopave.fhwa.dot.gov/Data/ClimateTool?mode=country.

FHWA. (2019). Freight Management and Operations - Freight Demand Modeling and Data Improvement: A Strategic Roadmap for Making Better Freight Investments. Retrieved from https://ops.fhwa.dot.gov/freight/freight_analysis/fdmdi/index.htm

Fu, X., Lam, W. H. K., & Meng, Q. (2014). Modelling Impacts of Adverse Weather Conditions on Activity-Travel Pattern Scheduling In Multi-Modal Transit Networks. *Transportmetrica B: Transport Dynamics, 2*(2), 151-167. doi:10.1080/21680566.2014.924084.

Hallenbeck, M., Rice, M., Smith, B. L., Cornell-Martinez, C., & Wilkinson, J. (1997). *Vehicle Volume Distributions by Classification*. Retrieved from https://trid.trb.org/view/665913.

Hanbali, R. M., & Kuemmel, D. A. (1993). Traffic Volume Reductions Due to Winter Storm Conditions. *Transportation Research Record,* (1387).

Hsiao, C. (2005). *Why Panel Data? University of Southern California*. IEPR working paper 05.33.

Hyun, K. K. (2016). *Network-Wide Truck Tracking Using Advanced Point Detector Data* Available from Dissertation Abstracts International. Retrieved from http://www.pqdtcn.com/thesisDetails/7A9A972180D7DAF737E6BB4299831BDC.

Ivanov, B., Xu, G., Buell, T., Moore, D., Austin, B., & Wang, Y. (2008). *Storm Related Closures Of I-5 And I-90 : Freight Transportation Economic Impact Assessment Report, Winter 2007-2008*. Retrieved from https://rosap.ntl.bts.gov/view/dot/17218.

Keay, K., & Simmonds, I. (2005). The Association of Rainfall and Other Weather Variables with Road Traffic Volume in Melbourne, Australia. *Accident Analysis and Prevention, 37*(1), 109-124. doi:10.1016/j.aap.2004.07.005.

Knapp, K. K., & Smithson, L. D. (2000). Winter Storm Event Volume Impact Analysis Using Multiple-Source Archived Monitoring Data. *Transportation Research Record: Journal of the Transportation Research Board, 1700*(1), 10-16. doi:10.3141/1700-03.

Kockelman, K. M. (1998). Changes in Flow-Density Relationship Due to Environmental, Vehicle, and Driver Characteristics. *Transportation Research Record: Journal of the Transportation Research Board, 1644*(1), 47-56. doi:10.3141/1644-06.

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. London: Chapman and Hall/CRC. doi:10.1201/9781420064254 Retrieved from http://www.crcnetbase.com/isbn/9781420064254.

Liu, Z., Li, X., Li, R., & Shang, P. (2015). Inclement Weather Impacts on Urban Traffic Conditions. *Cictp 2015* (pp. 2213-2227) doi:10.1061/9780784479292.206 Retrieved from http://ascelibrary.org/doi/abs/10.1061/9780784479292.206.

Maze, T. H., Agarwal, M., & Burchett, G. (2006). Whether Weather Matters to Traffic Demand, Traffic Safety, and Traffic Operations and Flow. *Transportation Research Record: Journal of the Transportation Research Board, 1948*(1), 170-176. doi:10.1177/0361198106194800119.

Maze, T. H., Crum, M. R., & Burchett, G. (2005). An Investigation of User Costs and Benefits of Winter Road Closures.

Melillo, J. M. (2014). *Climate Change Impacts in the United States.* (). Washington, DC: U.S. Global Change Research Program. Retrieved from https://doi.org/10.7930/J0Z31WJ2.

NOAA. (2015). Dew Point vs Humidity. Retrieved from https://www.weather.gov/arx/why_dewpoint_vs_humidity.

Pierce, D., & Short, J. (2012). Road Closures and Freight Diversion. *Transportation Research Record: Journal of the Transportation Research Board, 2269*(1), 51-57. doi:10.3141/2269-06.

Roh, H., Datla, S., & Sharma, S. (2013). Effect of Snow, Temperature and Their Interaction on Highway Truck Traffic. *Journal of Transportation Technologies, 3*(1), 24-38. doi:10.4236/jtts.2013.31003.

Roh, H., Sharma, S., & Datla, S. (2014). The Impact of Cold and Snow on Weekday and Weekend Highway Total and Passenger Cars Traffic Volumes. *The Open Transportation Journal, 8*(1), 62-72. doi:10.2174/1874447801408010062.

Roh, H., Sharma, S., Sahu, P. K., & Datla, S. (2015). Analysis and Modeling of Highway Truck Traffic Volume Variations during Severe Winter Weather Conditions in Canada. *Journal of Modern Transportation, 23*(3), 228-239. doi:10.1007/s40534-015-0082-2

Rowell, M., Gagliano, A., Wang, Z., Goodchild, A., Sage, J., & Jessup, E. (2012). *Improving Statewide Freight Routing Capabilities for Sub-National Commodity Flows*. Retrieved from https://rosap.ntl.bts.gov/view/dot/25079.

Sun, Z., & Ban, X. (. (2013). Vehicle Classification Using GPS Data. *Transportation Research Part C: Emerging Technologies, 37*, 102-117. doi:https://doi.org/10.1016/j.trc.2013.09.015.

Tessier, M., Morency, C., & Saunier, N. (2016). Impact of Weather Conditions on Traffic: Case Study of Montreal's Winter. Paper presented at the *95th Annual Meeting of the Transportation Research Board.*

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography, 46*, 234-240. doi:10.2307/143141.

Winston, C., & Shirley, C. (2004). The Impact of Congestion on Shippers' Inventory Costs. *Federal Highway Administration, Washington DC.*

Chapter 5

5 Applications

Two applications highlighting specific contributions of this work are presented in this chapter. The applications include the estimation of commercial vehicles' weight distribution on roads and the identification of the change in vehicle-miles-traveled (VMT) due to weather events. Through these applications, we suggest the ways in which the developed models can be used for policy analysis, travel demand forecasting, and operations.

5.1 Estimation of Truck Weight by Road Link

5.1.1 Introduction

Weight data of commercial trucks is a key component for freight modeling, pavement management, and pavement design. Particularly, the *Mechanistic-Empirical Pavement Design Guide (M-E PDG)* software requires site-specific, high-quality truck wheel load data as inputs (FHWA, 2019). However, this data is not widely available since it is typically only measured by Weigh-in-Motion (WIM) or static enforcement scales which are sparsely located along the highway network (Hernandez & Hyun, 2019). Further, weight data is not collected by vehicle detection stations (VDS), inductive loop detectors (ILD), or GPS based tracking methods (Hernandez, 2014). Hence, there is a need to identify another data source that can provide truck weight data for all road segments. In this application, we suggest a way in which our industry classification model can be used to address this critical data gap and compare our method to observed weight data gathered from 40 WIM sites in Arkansas.

5.1.2 Methods

The method can be divided into three steps: (1) identification of complete and fully connected truck paths from GPS data, and (2) application of industry classification model on truck GPS data, and (3) estimation of commodity tons on roads.

First, we used *path identification* algorithm of Chapter 1 (section 1.4.2), to identify fully connected complete truck paths from truck GPS data of Arkansas. Next, we applied the industry classification model of Chapter 3 on trucks and predicted the industry served. Afterward, we calculated the total number of trucks for each industry group on each road link. This calculated number of trucks was not that of the total truck population but a sample. Thus, we expanded the truck GPS sample to represent the entirety of the truck population. Expansion factors were derived through the comparison between the GPS volumes and truck traffic volumes measured by WIM sensors. On average, the statewide sample of GPS data in Arkansas represented 10-15% of the total truck traffic (Akter, Hernandez, Diaz, & Ngo, 2018; Corro, Akter, & Hernandez, 2019).

The estimation of commodity tons on roads followed three sequential steps (Figure 5.1). We multiplied the GPS truck volume by the expansion factors and calculate the total truck volume by the industry for each road link. Later, we used commodity-specific average payload factors (tons per truck) (see Table 5.2) from Arkansas' statewide travel demand model to calculate commodity tons. The payload factors include only fully loaded trucks (ARDOT, 2012). All commodity tons were totaled to get total tons on a road segment for a specific time period (Eq. 5.1).

$$W_i = \sum_{j=1}^{n} v_{ij} \times p_{ij} \qquad (5.1)$$

Where,

W_i = Total truck weights on road link i

v_{ij} = Truck volume for commodity group j on road link i

p_{ij} = Payload factors for commodity group j on road link i

n = number of commodity groups observed on road link i

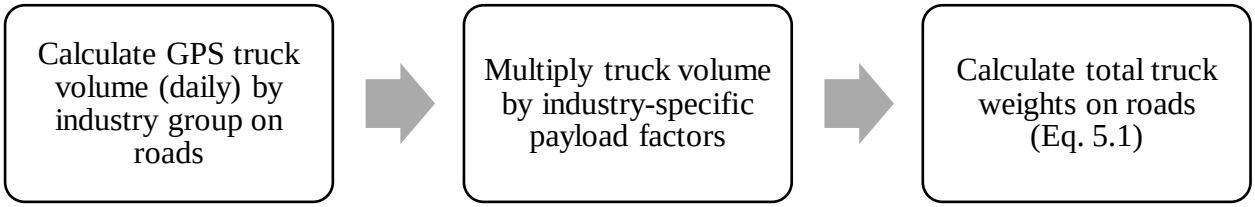


Figure 5.1 Sequential steps to calculate total truck weights on roads

We created the payload weight distribution of trucks on roads by plotting the calculated weight data on the x-axis and the percentage of trucks in that weight bin on the y-axis (Figure 5.2). Figure 5.2 shows an example of payload weight distribution on four road links near Lonoke, EL Dorado, Van Buren, and Cave City of Arkansas.

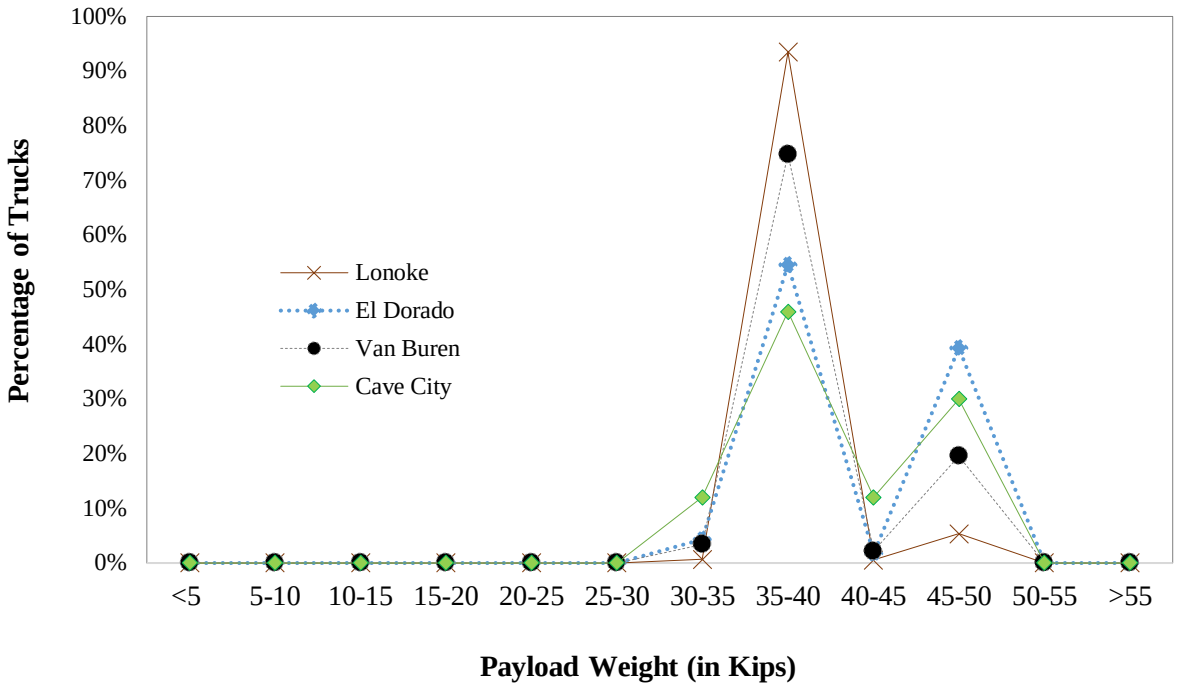


Figure 5.2 An example of payload weight distribution on road links

5.1.3 Discussion

For validation, we compared the predicted and observed truck weight estimates for 40 road segments in Arkansas (Figure 5.3). These 40 sites correspond to the locations of WIM sensor. The resulting highway daily truck volumes stratified by industry type show differing industry group proportions on each road link. For instance, we found that on Interstate 40 (I-40) around 94% of trucks were related to *manufacturing* industries while on Arkansas State Road ten (AR10) around 85% of trucks were related to *farm products* (Figure 5.3).

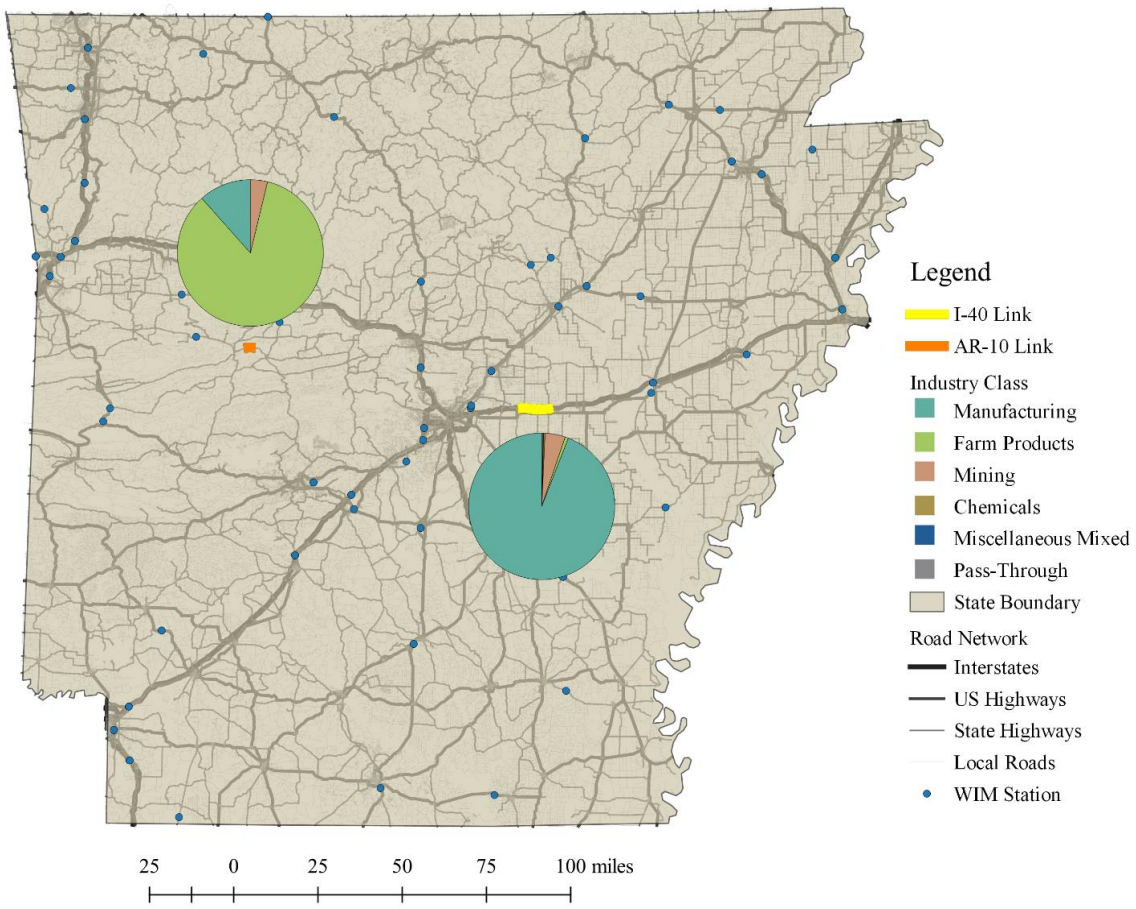


Figure 5.3 Distribution of industry class on roads

We calculated commodity weights from WIM data to compare with our predicted weights (Figure 5.4). In this comparison, we considered WIM data only for vehicles above FHWA Scheme F class 5. We assumed that empty truck weight varied from 10,000 – 26,000 lbs. based on their vehicle classes (FHWA, 2019). To understand the difference compared to the WIM data, we calculated the Absolute Percent Error (APE) between the two data sets using Eq. 5.2 (Table 5.1). The Mean Absolute Percent Error (MAPE) is calculated for each factoring method as follows:

$$APE_i = \left|\frac{w_i - g_i}{w_i}\right| \times 100\% \quad (5.2)$$

$$MAPE = APE/n \quad (5.3)$$

Where,

APE = Absolute Percent Error for station i

w_i = Total Commodity Weight from WIM for station i

g_i = Total Commodity Weight from GPS for road link near station i

$MAPE$ = Mean Absolute Percent Error

N = number of study sites (n = 40)

The APE ranges from 3% to 300% with MAPE of 71% across all 40 WIM sites. As the total daily truck volume increased in GPS data, the APE generally increases (Table 5.1). Since our payload factors were from fully loaded trucks, we expected this overestimation.

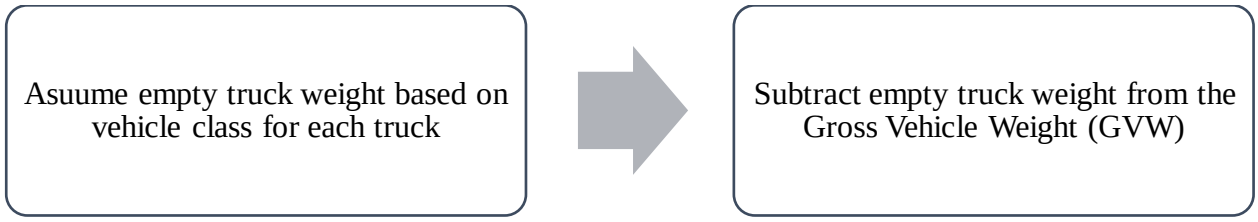


Figure 5.4 Steps to calculate commodity weights from WIM sensors

Since the APE and MAPE can only be used to gauge the general goodness of fit, we applied a statistical procedure, namely the *KS-test* to determine if the total truck weight estimated by our industry classification method and the WIM sensor were statistically similar (Table 5.1). Since truck weight distribution is not normally distributed (Hernandez & Hyun, 2019), we selected the *KS-test* over the *paired t-test*. The test is formulated as follows (Eq. 5.4 and 5.5):

Null Hypothesis: Total weights are not different between WIM data and GPS data

Alliterative Hypothesis: Total weights are different between WIM data and GPS data

In *KS-test* statistics,

$$D = x_i - y_i \quad \text{(Eq. 5.4)}$$

$$d = \frac{1.36}{\sqrt{n}} \quad \text{(Eq. 5.5)}$$

Where,

D = Maximum difference in our data

x_i = WIM weights at station i

y_i = GPS predicted weights on road near to the station i

d = Critical *KS-test* statistic

n = number of total stations (n = 40)

From our result, we found that-

The critical test statistic (d) = 0.22

Maximum difference in our data (D) = 0.07

Since the maximum difference in our data is smaller than critical test statistics (d), we can conclude that there is not enough evidence to say these two distributions are different. Hence, we can say that our predicted weight data showed a similar pattern to WIM weight data.

However, the relative (APE and MAPE) noted above can be partially attributed to inaccurate or misspecified truck payload factors. We assumed that using the average payload factors could be a source of error in our approach. To overcome this issue, we would need payload factors that reflect regional or site specific loading patterns. This would require new data to be collected, like via a travel survey like the national Vehicle Inventory and Use (VIUS) survey which was discontinued in 2007.

Further, in addition to five industry classes, our classification model identified *pass-through* trucks that did not have any industry association. Hence, we used average payload factors of all industry classes to calculate the total daily truck weight for *pass-through*. This could be another source of error in our prediction. However, in the future, we may consider applying an average payload factor to estimate the weight of pass-through trucks.

Table 5.1 Comparison of Total Commodity Weights

WIM Station Name	Total Commodity Weight from WIM (in Kips)	Total Commodity Weight from GPS (in Kips)	Absolute Percent Error (APE)
Lonoke	326,158	494,769	52%
Arkadelphia	312,580	358,397	15%
Glen Rose	280,113	365,777	31%
Texarkana	256,587	380,269	48%
Gilmore	230,755	138,381	40%
Lamar	150,343	267,890	78%
Dora	129,097	4,598	96%
Fayetteville	94,330	247,389	162%
Alma	80,366	258,090	221%
Texarkana	54,073	36,074	33%
Fort Smith	53,367	155,591	192%
Pine Bluff	41,224	83,106	102%
Fouke	40,081	19,838	51%
Bald Knob	40,045	32,194	20%
Rixey	38,963	75,951	95%
Jonesboro	29,345	25,981	11%
El Dorado	27,300	18,529	32%
Grady	23,123	15,237	34%
Van Buren	20,408	21,973	8%
Thornton	19,703	30,839	57%
Omaha	18,871	226	99%
Light	15,696	8,429	46%
Damascus	14,706	20,466	39%
Pindall	11,755	7,997	32%
Needmore	10,060	16,219	61%
Malvern	9,258	12,577	36%
Dardanelle	9,032	7,221	20%
Patterson	8,543	10,703	25%
Hot Springs	7,621	30,458	300%
Sunnydale	7,292	8,043	10%
Monette	6,676	7,463	12%
Bradley	5,562	514	91%
Bryant	5,328	2,710	49%
Searcy	5,091	9,027	77%
St. Charles	4,192	3,060	27%
Brinkley	3,282	1,504	54%
Berryville	2,669	6,675	150%
Cave City	2,349	993	58%
Monticello	2,011	824	59%
Pangburn	1,847	6,032	227%
Median APE			**49%**
MAPE			**71%**

Although our proposed approach does not exactly replicate WIM measured weights, it does allow for the estimation of weight at sites without WIM stations. Our approach can predict total daily truck weight and the distribution of that weight for any road link where GPS data are present. For instance, we calculated the daily average trucks weight on AR-10 from GPS data using this approach while no weight data were available from the WIM system (Table 5.2).

Table 5.2 Daily Truck Weights on AR-10 Road Link

Industry Class	Payload Factors *(Tons/Truck)	Daily Truck Volume	Daily Total Truck Weight (in Tons)
Manufacturing	19.08	43	821
Farm Products	16.26	314	5,105
Mining	23.96	14	335
Chemicals	20.67	0	0
Miscellaneous Mixed	21.66	0	0
Pass-through	20.34	0	0
Total		**371**	**6,262**

* 1 Ton = 2 Kips

This approach can be used to determine truck loads across the whole road network and thus, can assist in comprehensive pavement management. Further, this approach can help identify critical commodity-based freight corridors which can potentially lead to the development of commodity specific performance measures. Identification of commodity-based critical freight corridors is crucial for transportation planning agencies to prioritize their projects based on freight market value as well as volume and weight. Ultimately, this is an approach that can be used to fill out the data gap in the weight distribution on roads.

5.2 Effects of Weather Events on Vehicle-Miles-Traveled (VMT)

5.2.1 Introduction

Adverse weather events such as floods, heavy rainfall, storm, snowfall, and extreme heat can have major effects on traffic volumes (Melillo, 2014). While drivers of passenger vehicles may choose not to travel during inclement weather, freight truck drivers adhere to delivery schedules requiring them to alter their route rather than cancel a trip (Datla, Sahu, Roh, & Sharma, 2013). To assist freight trucks in rerouting and ensure efficient movement during adverse weather conditions, it is necessary for the state planning agencies to understand the effect of weather events. With the aim of identifying the effects of weather events on truck traffic, we applied our spatial regression model on truck GPS data to predict the change in Vehicle-Miles-Traveled (VMT) resulting from weather events. Since VMT is a combination of volume and miles, it captures the rerouting behavior of trucks more accurately than the only volume. This application of the spatial regression model can assist state and regional transportation agencies in developing freight-oriented programs and policies for winter maintenance and alternate route planning. Also, to assist the trucking industry to better plan accurate routes to estimate arrival times and revenue miles.

5.2.2 Methods

We used truck GPS and weather data as the primary inputs to develop a spatial regression model in this application. The method can be divided into three segments: (1) identification of complete and fully connected truck paths from GPS data, (2) calculation of VMT, and (3) estimation of a spatial regression model.

Using *path identification* algorithm of Chapter 1 (section 1.4.2), first, we identified complete truck paths and volumes on roads from GPS data. Later, the identified trucks' trip

lengths (in miles) and volumes were used to calculate the daily Vehicle Miles Traveled (VMT) for a specific road segment (Eq. 5.6). Next, the change in daily VMT was calculated by comparing the daily VMT to the average VMT over the year.

$$VMT_i = V_i \times MT_i \quad (5.6)$$

Where,

VMT_i = Vehicle Miles Traveled for road link *i*

V_i = Daily truck volume on road link *i*

MT_i = Daily trip length of trucks on road link *i*

Further, we collected daily weather variables like temperature, precipitation, and wind speed from the Modern-Era Retrospective analysis for Research and Applications (MERRA) and adverse weather events data such as flood, snowfall, storm, and drought for specific days from National Oceanic and Atmospheric Administration (NOAA). Finally, we estimated the Spatial Autoregressive (SAR) model using the change in VMT as the dependent variable and weather parameters as independent variables (Eq. 5.7) (Belotti, Hughes, & Mortari, 2017).

$$y_{it} = \rho W_{ij} y_{jt} + X_{it}\beta + u_i + \varepsilon_{it} \quad (5.7)$$

Where,

ρ = spatial autoregressive parameter

W_{ij} = An element of a spatial weights matrix W describing the spatial arrangement of the units in the sample. It is assumed that W is a pre-specified non-negative matrix of order N.

i = an index for the cross-sectional dimension (roads)

t = an index for the time dimensions (days)

$\boldsymbol{y_{it}}$ = Change in VMT on road *i* for day *t*

$\boldsymbol{X_{it}}$ = A vector of explanatory variables (weather parameters) of road *i* for day *t*

$\boldsymbol{\beta}$ = The coefficient of explanatory variables $\boldsymbol{X_{it}}$

u_i = a spatial specific effect; The standard reasoning behind spatial specific effects is that they control for all space-specific time-invariant variables whose omission could bias the estimates in a typical cross-sectional study. It is assumed that $\mu \sim N(0, \sigma_u^2)$ in the random-effects case, while the μ is a vector of parameters to be estimated in the fixed-effects variant.

ε_{it} = is an independently and identically distributed error term for road *i* On day *t* with zero mean and variance σ^2

5.2.3 Results

A comparison between ordinary least square regression (OLS) and SAR models shows that OLS model cannot capture the effects of the spatially dispersed variables since it does not consider the spatial autocorrelation of the dependent variable (Table 5.3). Unlike the OLS model, the developed SAR model shows that spatial autoregressive parameter rho (ρ) is positive (0.72) and statistically significant at the 99% level of confidence, which is evidence that truck vehicle miles traveled (VMT) are spatially autocorrelated. In other words, change in VMT on one road segment would affect the change in VMT on the neighboring road segments.

The positive coefficient values of Table 5.3 indicate an increase in daily VMT while the negative value indicates a decrease. Since the SAR model considers the spatial autocorrelation of the dependent variable, it can capture the effects of the spatially dispersed independent variables. For instance, the SAR model identified that if one road segment had an average VMT of 100 vehicle-miles and that link observed 1 mm snowfall, the VMT of that road segment would be

reduced to 95 vehicle-miles for that specific day. The lower AIC (Akaike information criterion) value of the SAR model also indicates a better predictive power of this model compared to the OLS model in capturing the effect of weather variables (Table 5.3).

Table 5.3 Comparison between OLS and SAR Models for VMT

Independent Variables	Ordinary Least Square Regression (OLS)	Spatial Autoregressive Models (SAR)
Snowfall		-0.05***
Storm Events	-0.17***	-0.10***
Extreme Heat	-0.07***	-0.03***
Weekday	0.68***	0.29***
Spring	0.05***	0.02***
Summer	0.13***	0.04***
Fall	0.10***	0.03***
Constant	0.51***	0.14***
Spatial, rho (ρ)		0.72***
R-squared	**0.54**	**0.54**
AIC	**829.5**	**214.6**

***significant at 99% confidence level; **significant at 95% confidence level; *significant at 90% confidence level

5.2.4 Conclusion

Since the weather impacts to or in the vicinity of Primary Freight Network (PFN) segments have far reaching effects on freight movements across the nation, it is necessary to identify the effects accurately (Winston & Shirley, 2004). To capture the effects during adverse weather events, the state planning agencies can use our model that estimates the change in VMT on roads. They can better understand alternate route usage, and plan deicing strategies on the primary and alternate routes as well as change signalization operations to minimize increases in traffic along arterial routes during storms, for example. Moreover, impacts of weather events such as rerouting and displaced congestion cause shipment delays, depreciation of goods, and inventory holding costs (Winston & Shirley, 2004). Thus, it is crucial for the trucking industry to understand the change in routes during adverse weather conditions. For the trucking industry,

understanding the change in VMT will help to better estimate route miles when inclement weather is predicted during the shipment. This will help shippers to accurately calculate revenue miles. The main contribution of this application is capturing the spatial effect of weather variables on truck volume and trip length simultaneously. VMT better captures the effects of weather on rerouting or temporal delays to trips. Ultimately, the use of GPS allows us to measure the changes in VMT at dispersed locations unlike static traffic data collection sites such as Weigh-In-Motion (WIM) or AADT

5.3 References

Akter, T., Hernandez, S., Diaz, K. C., & Ngo, C. (2018). Leveraging Open-Source GIS Tools to Determine Freight Activity Patterns from Anonymous GPS Data. Paper presented at the *2018 AASHTO GIS for Transportation Symposium.*

ARDOT. (2012). *Arkansas Statewide Travel Demand Model*.

Belotti, F., Hughes, G., & Mortari, A. P. (2017). Spatial Panel-Data Models Using Stata. *The Stata Journal, 17*(1), 139-180. doi:10.1177/1536867X1701700109.

Corro, K. D., Akter, T., & Hernandez, S. (2019). Comparison of Overnight Truck Parking Counts with GPS-Derived Counts for Truck Parking Facility Utilization Analysis. *Transportation Research Record, 2673*(8), 377-387. doi:10.1177/0361198119843851.

Datla, S., Sahu, P., Roh, H., & Sharma, S. (2013). A Comprehensive Analysis of The Association of Highway Traffic with Winter Weather Conditions. *Procedia - Social and Behavioral Sciences, 104*, 497-506. doi:10.1016/j.sbspro.2013.11.143.

FHWA. (2019). WIM Data Analyst's Manual. Retrieved from https://www.fhwa.dot.gov/pavement/wim/pubs/if10018/if10018.pdf.

Hernandez, S. V. (2014). *Integration of Weigh-in-Motion and Inductive Signature Data for Truck Body Classification* Available from Dissertations & Theses Europe Full Text: Science & Technology. Retrieved from https://search.proquest.com/docview/1648677216.

Hernandez, S., & Hyun, K. (. (2019). Fusion of Weigh-In-Motion and Global Positioning System Data to Estimate Truck Weight Distributions at Traffic Count Sites. *Journal of Intelligent Transportation Systems, ,* 1-15. doi:10.1080/15472450.2019.1659793.

Melillo, J. M. (2014). *Climate Change Impacts In The United States*. Washington, DC: U.S. Global Change Research Program. Retrieved from https://doi.org/10.7930/J0Z31WJ2.

Winston, C., & Shirley, C. (2004). The Impact of Congestion on Shippers' Inventory Costs. *Federal Highway Administration, Washington DC.*

Conclusion

The method presented in this dissertation integrating a large stream of anonymous mobile sensor data and advanced machine learning techniques will uniquely fill the existing research gaps in freight transportation studies. The resulting four models developed to extract industry-specific truck activity patterns and weather-related rerouting behaviors of trucks address the critical methodological needs for freight planning and operation applications.

First, a Multinomial Logistic (MNL) regression model was developed to identify the key operational characteristics of freight that define the carried commodities of trucks. To develop this model, we applied three sets of heuristic algorithms: *stop identification*, *path identification*, and *trip identification.* These algorithms extracted stop time of day, stop location, stop duration, stop coverage, truck paths, trip length, and trip duration from a large stream of anonymous truck GPS data. The model identified stop time of day, stop duration, and trip lengths as the statistically significant features that change over industry types. Although the developed MNL model identified commodity-specific operational characteristics, the log likelihood suggests the need for using advanced machine learning techniques to capture unexplained variabilities in the data (Caruana and Niculescu-Mizil, 2006).

Using the salient features of MNL model, next we developed a *K*-means clustering model to extract representative freight activity patterns that can support and validate activity-based models. *K*-means clustering is an unsupervised technique of machine learning that can derive patterns from anonymous data. Our clustering model was developed using approximately 300,000 daily truck movement records. It extracted six unique and representative activity patterns that can be used to support and validate activity-based models. However, due to the anonymity of GPS data, it was not possible to directly "observe" the demographic characteristics

of the trucks within each representative pattern. Hence, there was still a need to apply supervised machine learning techniques to predict industry-served or commodity-carried of freight trucks from operational characteristics.

In response to that methodological need, we applied the random forest algorithm, a supervised machine learning tool, to develop an industry classification model using anonymous truck GPS data. The operational characteristics derived from heuristics algorithms were adapted to create 11-element feature vectors. Next, a proximity analysis was conducted to create the probability matrix of the industry class. A total of 31 business categories were added to the probability matrix. Then, aerial imageries were used to make 2,064 *groundtruth* data with industry class labels. Finally, we developed the industry classification model by splitting the *groundtruth* data into a 66/34 training/testing set. Our developed model can predict the carried commodity of trucks with 90% accuracy and 0.97 ROC area. The model was developed in a way so that it can discern the industry class of a truck while maintaining the anonymity of the data. For instance, the model predicts the industry class of a commercial truck from its operational characteristics but does not disclose any identifiable information such as driver's name, fleet, or company's private information. *Manufacturing, farm products, mining, chemicals, miscellaneous mixed,* and *pass-through* are six industry classes that can be predicted from the model. To validate the classification model, we applied it to 300,000 daily truck movements of Arkansas and compared the result with the Arkansas Statewide Travel Demand Model (AR STDM). The comparison reveals a commodity flow pattern similar to AR STDM. Therefore, we suggest that the classification model can be used to support and validate commodity-based freight forecasting models. However, there is scope to improve the performance of our industry classification

model. We suggest three solutions including increasing the training data sample size for the minority classes, changing the buffer distance, and disaggregating the six industry classes.

State transportation planning agencies and freight industry both strive to understand the rerouting behavior of commercial trucks during adverse weather events. To address this critical research need, we developed a spatio-temporal regression model fusing fixed sensors (e.g., WIM) with weather data. The developed model identified how one-unit change in weather related variables (i.e., snow mass, humidity, and surface runoff) could affect daily truck traffic volume of a route and its neighboring routes. In essence, it captured the rerouting behavior of trucks. We used historical truck volume, computed over a five-year period (2011-2015) to predict future truck volume (2016). The model predicts both spatial and temporal effects of weather on truck traffic volumes and hence, can be used to support and improve long-range transportation planning as well as maintenance operations. The model also can help trucking industries to estimate billable miles more accurately.

Although the four developed models have several freight planning applications, we described two applications in this dissertation. Our industry classification model can be used to estimate commercial vehicles' weight distribution on roads and identify the change in vehicle-miles-traveled (VMT) due to weather events. The estimation of commercial trucks' weight distribution on roads can support transportation engineers in the pavement management and design. The determination of change in vehicle miles traveled (VMT) due to weather events can better capture the route changes. Further, the models can also be used to identify commodity-specific critical freight corridors that are necessary for prioritizing freight projects.

Our developed choice model identified the significant operational characteristics that change based on the carried commodity of trucks. These statistically significant operational

characteristics were used to develop our clustering model that can identify the unique daily activity patterns of freight trucks. These activity patterns can be used to support and validate activity based travel demand models. Additionally, our classification model demonstrated that operational characteristics of trucks including the number of stops, stop location, stop duration, stop time of day, trip length, and trip duration have distinct patterns based on commodity carried and industry served. Finally, our spatio-temporal model identified the effects of weather events on truck traffic. This model captured the rerouting behaviors of freight trucks that can be used to support and improve long-range transportation planning as well as maintenance operations. Ultimately, being a combination of four predictive models, this dissertation can support State or Federal agencies for policy analysis, travel demand forecasting, and operations.

Reference

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. Paper presented at the 161-168. doi:10.1145/1143844.1143865 Retrieved from http://dl.acm.org/citation.cfm?id=1143865.